Labels Generated by Large Language Models Help Measure People's Empathy *in Vitro*

Md Rakibul Hasan, *Graduate Student Member, IEEE,* Yue Yao, *Graduate Student Member, IEEE,* Md Zakir Hossain, *Member, IEEE,* Aneesh Krishna, Imre Rudas, Shafin Rahman, and Tom Gedeon, *Senior Member, IEEE*

***Abstract*—Large language models (LLMs) have revolutionised many fields, with LLM-as-a-service (LLMSaaS) offering accessible, general-purpose solutions without costly task-specific training. In contrast to the widely studied prompt engineering for directly solving tasks (*in vivo*), this paper explores LLMs' potential for *in-vitro* applications: using LLM-generated labels to improve supervised training of mainstream models. We examine two strategies – (1) noisy label correction and (2) training data augmentation – in empathy computing, an emerging task to predict psychology-based questionnaire outcomes from inputs like textual narratives. Crowdsourced datasets in this domain often suffer from noisy labels that misrepresent underlying empathy. We show that replacing or supplementing these crowdsourced labels with LLM-generated labels, developed using psychology-based scale-aware prompts, achieves statistically significant accuracy improvements. Notably, the RoBERTa pre-trained language model (PLM) trained with noise-reduced labels yields a state-of-the-art Pearson correlation coefficient of 0.648 on the public NewsEmp benchmarks. This paper further analyses evaluation metric selection and demographic biases to help guide the future development of more equitable empathy computing models. Code and LLM-generated labels are available at https://github.com/hasan-rakibul/LLMPathy.**

***Index Terms*—Empathy detection, Large language model, Natural language processing, Label noise, NewsEmp.**

I. Introduction

Large language models (LLMs) have become a go-to approach across a variety of tasks, such as emotion recognition [1] and empathy detection [2], [3]. Due to high computational demands, coupled with environmental impact, training or fine-tuning LLMs often becomes costly. This limitation has led to increasing adoption of LLMs as a service (LLMSaaS), where users access trained LLMs via online APIs with computation on cloud [4]. LLMSaaS can be utilised *in-vivo*, *i.e.*, prompt engineering to directly solve tasks such as named entity recognition [5], sentiment analysis [6] and empathy detection [2], [3], or *in-vitro* [7][1], *i.e.*, integrating LLM outputs into *other* models.

We are motivated by the following considerations. First, most current applications of LLMSaaS leverage LLM outputs *in-vivo* [2], [3], [5], [6]. We shift to their utility *in-vitro* to fine-tune smaller pre-trained language models (PLMs)[2] like RoBERTa [8]. In particular, we propose to utilise LLMSaaS in a *data-centric AI* approach [9] to **(1)** enhance the quality of training labels and to **(2)** increase the amount of quality training data for supervised training of PLMs.

Second, for representation learning, maintaining data quality is critical – captured succinctly by the phrase, "garbage in, garbage out" [10]. While deep learning research has mostly focused on proposing new algorithms, improvement in data-centric AI is equally important [9]. As a data-centric AI approach, we leverage LLMs to enhance data quality. The effectiveness of our proposed approach is demonstrated in an emerging field – empathy detection.

Empathy is defined as *"an affective response more appropriate to another's situation than one's own"* [11]. In psychology, various questionnaires have been developed to measure empathy. Empathy computing[3], in computer science, complements these psychology-based methods by aiming to map the questionnaire outcomes from input stimuli such as textual narratives, audiovisual interactions and physiological signals [12]. One well-known questionnaire is the empathy measurement scale proposed by Batson *et al.* [13], which assesses empathy across six dimensions: sympathetic, moved, compassionate, tender, warm and soft-hearted.

Empathy *computing* offers the potential to improve people's empathic skills, which in turn strengthens interpersonal relationships across various human interactions [12]. In healthcare, for example, empathic writing in medical documents (*e.g.*, patient reports) can promote understanding and trust between clinicians and patients [15]. Similarly, in education, written communication like emails and feedback on assignments has become a vital medium for expressing care and addressing students' emotional needs [16]. Journalism also demonstrates the importance of empathy in written narratives. For example,

M R Hasan, Y Yao, M Z Hossain, A Krishna and T Gedeon are with School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley WA 6102, Australia.

I Rudas is with Obuda University, Budapest, Hungary.

S Rahman is with North South University, Dhaka 1229, Bangladesh.

M R Hasan is also with BRAC University, Dhaka 1212, Bangladesh.

Y Yao, M Z Hossain and T Gedeon are also with The Australian National University, Canberra ACT 2600, Australia.

T Gedeon is also with Obuda University, Budapest, Hungary.

E-mail: {Rakibul.Hasan, Zakir.Hossain1, A.Krishna, Tom.Gedeon}@curtin.edu.au, yue.yao@anu.edu.au, rudas@uni-obuda.hu, shafin.rahman@northsouth.edu

Corresponding author: M R Hasan

[1]Like [7], we use the term *"in-vitro"* to refer to leveraging LLM outputs out of the box in a different model.

[2]We use *"pre-trained language models (PLMs)"* to refer specifically to smaller models like the BERT family of models, distinguishing them from LLMs, which are also pre-trained but significantly larger.

[3]We use the terms empathy computing, detection, prediction and measurement interchangeably.

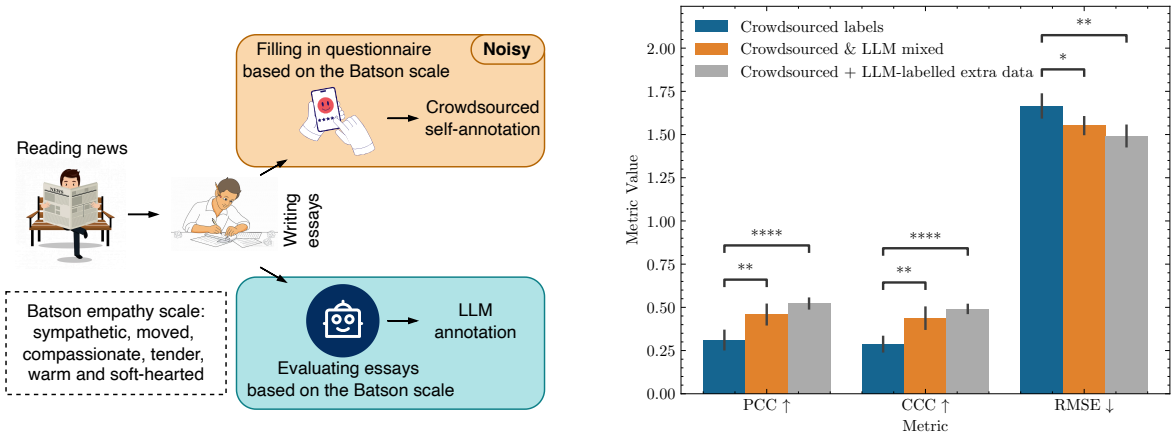

Fig. 1. **Left:** Overview of traditional and LLM-based methods to annotate essays for detecting empathy in the essays written in response to news articles. To be shown in our experiment, existing crowdsourced self-annotation through questionnaires is found to be incorrect in many samples. Our proposed approach involves annotating the essays using LLM, which is then used to reduce label noise and to get additional training data. **Right:** Impact of our LLM usage is showcased through a performance comparison between existing crowdsourced annotations, LLM-based label noise correction and the inclusion of additional data labelled by LLM. Statistical significance is calculated using `Statannotations` package [14], where * means $0.01 < p\text{-value} \leq 0.05$, ** means $0.001 < p\text{-value} \leq 0.01$ and **** means $p\text{-value} \leq 0.0001$ (*I.e.* more * means higher statistical significance).

a news article on a family's recovery after a devastating event often goes beyond factual reporting and offers a compassionate perspective that engages readers emotionally and deepens their connection to the news story. Specifically, this paper measures people's empathy in essays written in response to scenarios reported in newspaper articles.

Empathy is inherently subjective, and machine learning models, including LLMs, used for its detection exhibit biases across different demographic groups [17]. We explore potential sources of such biases (Section IV-E). Additionally, while the Pearson correlation coefficient (PCC) remains the most commonly used evaluation metric in empathy computing [12], it does not account for the magnitude of the error. To address this, we advocate for adopting the concordance correlation coefficient (CCC) (Section IV-B).

Neural networks are prone to memorising training data, *i.e.*, overfitting. This issue is exacerbated by noisy labels, where traditional regularisation techniques like dropout and weight decay often fall short [18]. A major challenge in ensuring data quality is, therefore, addressing label noise, defined as labels that deviate from their intended values. It is a significant challenge in empathy computing datasets collected through crowdsourcing. Platforms like Amazon Mechanical Turk offer quick access to large participant pools. Accordingly, crowdsourcing with questionnaire-based self-assessment labelling is a popular way of collecting data in computational social psychology and human behaviour research, such as empathy [19] and emotion recognition [20]. However, such data often suffer from inaccuracies due to inattentiveness or multitasking among participants, compromising data reliability [21]–[24]. This necessitates strategies to enhance data quality post-collection.

The overarching goal of this paper is to address the question: *"How can LLMs enhance training of PLMs to improve empathy computing accuracy?"* As illustrated in Fig. 1, our proposed *in-vitro* applications achieve statistically significant performance improvements. The first application, which automatically adjusts training labels, demonstrates consistent performance improvement across all metrics compared to the baseline PLM trained on the original dataset. The second application, leveraging additional LLM-labelled training data, further enhances model performance, yielding the highest statistically significant performance gains with a p-value < 0.0001 [14].

Our key contributions are summarised as follows:

1) We propose two *in-vitro* applications of LLMs: mitigating label noise and getting additional training data for PLMs.
2) We design a novel scale-aware prompt that enables LLMs to annotate data while adhering to annotation protocols grounded in theoretical frameworks.
3) We investigate challenges in empathy computing datasets and advocate for new evaluation metrics.
4) Our proposed methods achieve statistically significant performance improvements over the baseline models across multiple datasets and set a new state-of-the-art empathy computing performance.

II. Related Work

A. LLM in Data Annotation

The advent of LLMs has inspired numerous studies exploring LLMs' application in data annotation, often positioning them as a substitute for traditional human annotation. For instance, Niu *et al.* [25] examined the potential of LLMs in emotion annotation tasks and reported that LLMs can generate emotion labels closely aligned with human annotations. Similarly, Wang *et al.* [26] explored the utility of LLMs in annotating datasets for various natural language processing tasks, including sentiment analysis, question generation and topic classification. While they highlighted the cost-effectiveness of LLM-based annotations, they also noted LLMs' limitations compared to human annotators. Departing from this line of work, our approach explores LLM-generated labels to enhance the training of PLMs. Specifically, we

integrate LLM-generated labels with human-generated labels rather than exclusive use of either LLM- or human-generated labels.

We examine our approach in two distinct applications: label adjustment and training data enhancement. Related to our first application, Hasan *et al.* [27] also explored label noise adjustment, but that approach relies on subjects' demographic information (*e.g.*, age, gender and race) in the prompting process. In contrast, our method deliberately avoids any use of demographic details to mitigate potential biases inherent in LLM training. Additionally, the reliance on demographic information may not always be feasible, which makes our approach more broadly applicable. Another key difference with Hasan *et al.* [27]'s prompting strategy is the use of multiple input-output examples: while they rely on few-shot prompting with multiple example pairs to elicit LLM output in a consistent style, our approach does not require such examples, yet still achieves consistent outputs. Furthermore, they experimented solely on the GPT-3.5 LLM, whereas we explore both Llama 3 70B [28] and GPT-4 [29] LLMs in a recent dataset.

B. Learning with Label Noise

Noise-robust learning has been extensively studied in classification settings, especially for computer vision, leaving textual regression tasks under-explored [30]. Such learning algorithms were demonstrated either explicitly through dedicated methods [31]–[34] or implicitly as part of broader semi-supervised learning frameworks [35], [36]. Dedicated methods such as [33], [34] address noise by a de-noising loss function based on the usual cross-entropy loss function in classification. Semi-supervised methods [35], [36] generate pseudo-labels based on class probabilities produced by `sigmoid` or `softmax` activations. Both categories of methods have shown strong performance in classification tasks; however, they are not directly applicable to regression, where the target space is continuous and lacks discrete output logits. Our *in-vitro* approach operates natively in the regression setting with the usual regression loss function and does not require any conversion to pseudo-class probabilities.

The study of label noise in textual regression remains limited. Wang *et al.* [30]'s approach iteratively identifies noisy examples and applies one of three strategies: discarding noisy data points, substituting noisy labels with pseudo-labels, or resampling clean instances to balance the dataset. While effective in identifying extreme outliers, Wang *et al.* [30] stated that their approach struggles in detecting mild disagreements. They further noted that their method performs worse in general-purpose datasets, compared to knowledge-dense domains, such as clinical notes and academic papers. Our approach leverages LLMs as an external "teacher" to directly correct noisy labels in a *single* pass. Given the general-purpose nature of LLMs, our approach holds the potential to be effective across different domains.

C. Empathy Computing

Empathy computing is an emerging field, with significant advancements in textual empathy prediction [12]. For a detailed overview of its progress, we refer to a recent systematic literature review by Hasan *et al.* [12].

In textual empathy computing, the most widely studied context is detecting people's empathy in response to newspaper articles. The Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA) Shared Tasks (2021–2024) have spurred various approaches leveraging PLMs on this task. Most approaches predominantly employed fine-tuning PLMs, with RoBERTa being the most preferred PLM [37]–[49]. Some studies have explored other BERT-based PLMs [50]–[53] or ensemble strategies combining multiple PLMs [54]–[56]. The suitability of fine-tuning RoBERTa is further validated by Qian *et al.* [38], who reported that simple fine-tuning of RoBERTa outperformed more complex multi-task learning in textual empathy computing. Overall, fine-tuning PLMs has emerged as the predominant approach for this task, with RoBERTa being the leading model [12].

More recently, LLMs have been explored for textual empathy prediction through rephrasing text for data augmentation [27], [48], fine-tuning [2] and prompt engineering [3]. Hasan *et al.* [27] adds multi-layer perception layers on top of a RoBERTa PLM to process demographic data, while Li *et al.* [2]'s fine-tuning of LLM demands significant computational resources. Unlike these methods, our approach leverages LLM-generated labels to enhance fine-tuning of a standard RoBERTa PLM, without demographic data or high computational costs.

III. Method

A. Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ represent a dataset, where x_i is the i-th input text sequence, and $y_i \in \mathbb{R}$ denotes corresponding continuous empathy score. Empathy, being a psychological construct, is challenging to annotate due to subjectivity. Consequently, the target variable y_i often suffers from noise, which is particularly significant in crowdsourced annotations (refer to Section IV-D1 for evidence of noise). We denote the noisy ground-truth empathy score as $\tilde{y}_i$, which serves as a proxy for the true, unobserved empathy score y_i. Thus, the dataset can be reformulated as $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$. Our goal is to develop a model $\mathcal{F} : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is the space of text sequences, such that $\mathcal{F}(x_i)$ accurately estimates y_i.

The dataset $\mathcal{D}$ is randomly partitioned into three non-overlapping subsets: a training set $\mathcal{D}_{\text{train}} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N_{\text{train}}}$, used to train the models; a validation set $\mathcal{D}_{\text{val}} = \{(x_j, \tilde{y}_j)\}_{j=1}^{N_{\text{val}}}$, used for optimising the training and tuning hyperparameters; and a hold-out test set $\mathcal{D}_{\text{test}} = \{(x_k, \tilde{y}_k)\}_{k=1}^{N_{\text{test}}}$, reserved for final model evaluation. The reserved D_{test} has not been altered in any way through the experiments.

Large language models (LLMs) can be leveraged to improve training in such noisy scenarios, specifically to assist a smaller pre-trained language model (PLM) in better estimating the empathy score. We consider two *in-vitro* approaches for leveraging LLM-generated labels to enhance model training. While such LLM use incurs non-trivial computational or API costs, it is a one-time expense during data preprocessing.

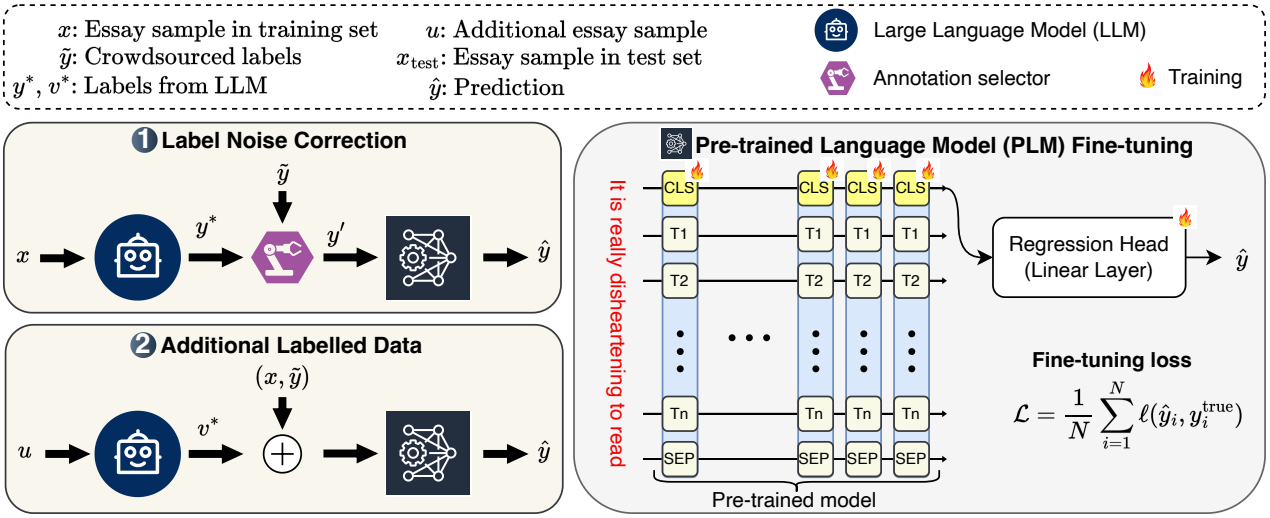

Fig. 2. Overview of our proposed *in-vitro* applications of large language models (LLMs) for enhancing textual empathy prediction with pre-trained language models (PLMs). Application 1 involves correcting noisy labels in an existing dataset using an LLM. Application 2 utilises an LLM to label additional text data, which is then added to the existing training dataset.

TABLE I
PLAIN VS SCALE-AWARE PROMPT TEMPLATES TO GENERATE LABELS FROM LLM.

Scheme	System prompt	User prompt
Plain	Your task is to measure the empathy of individuals based on their written essays. Human subjects wrote these essays after reading a newspaper article involving harm to individuals, groups of people, nature, etc. The essay is provided to you within triple backticks.	Essay: ```{essay}``` Now, provide empathy score between 1.0 and 7.0, where a score of 1.0 means the lowest empathy, and a score of 7.0 means the highest empathy. You must not provide any other outputs apart from the scores
Scale-aware	Your task is to measure the empathy of individuals based on their written essays. You will assess empathy using Batson's definition, which specifically measures how the subject is feeling each of the following six emotions: sympathetic, moved, compassionate, tender, warm and softhearted. Human subjects wrote these essays after reading a newspaper article involving harm to individuals, groups of people, nature, etc. The essay is provided to you within triple backticks.	Essay: ```{essay}``` Now, provide scores with respect to Batson's empathy scale. That is, provide scores between 1.0 and 7.0 for each of the following emotions: sympathetic, moved, compassionate, tender, warm and softhearted. You must provide comma-separated floating point scores, where a score of 1.0 means the individual is not feeling the emotion at all, and a score of 7.0 means the individual is extremely feeling the emotion. You must not provide any other outputs apart from the scores.

All subsequent training and inference are performed using smaller PLMs, which keeps the overall resource requirement comparable to standard textual regression workflows.

B. Applications of Large Language Model in-Vitro

Our proposed framework leverages LLM for empathy prediction, as illustrated in Fig. 2. The first application reduces label noise, while the second application increases the amount of training data by incorporating additional labelled data from LLM. Improved training data from these two applications is fed to a pre-trained language model (PLM) for final empathy prediction.

1) Prompt Design: To generate labels from LLM, one could instruct it to directly output the label (referred to as the *plain* prompting scheme). Instead, we design a *scale-aware* scheme, which includes subscales of the original labelling protocol. As presented in Table I, each prompting scheme is structured into two primary components: the system prompt, which defines the task of the LLM and establishes the expected behaviour, and the user prompt, which contains specific input texts we want to annotate and the range of the outputs.

The motivation behind leveraging a scale-aware scheme includes natural alignment with the Psychology-grounded labelling protocol that was used for crowdsourced raters in the original study. For example, the NewsEmp datasets used Batson's Empathy scale, which has six subscales, and so our scale-aware prompt instructs the LLM to provide scores on the subscales (Table I). The difference between the human- and LLM-generated labelling is minimal, with the only difference being who labels the data. Therefore, following the same protocol designed for crowdsourced raters, LLM outputs across the subscales are averaged to calculate a single empathy score y^*. The following subsections present details about how we use these LLM labels in empathy computing.

2) Application 1: Noise Mitigation in Labels: The LLM-generated labels y^* are used to identify and replace potentially noisy samples in the original crowdsourced annotation $\tilde{y}$. Noisy samples are identified based on the difference between $\tilde{y}$ and y^*. Like [27], a revised label y'_i is defined as:

$$y'_i = \begin{cases} y_i^* & \text{if } |\tilde{y}_i - y_i^*| > \alpha \\ \tilde{y}_i & \text{otherwise,} \end{cases} \tag{1}$$

where α is a predefined threshold, referred to as the *annotation selection threshold*, which determines which label to use for which sample.

This selection threshold can be any real number between 0 and the range of the empathy score (*i.e.*, $7-1=6$ for the NewsEmp dataset). A smaller α results in a more aggressive correction (replacing labels even for small differences). This could, however, lead to a larger distribution mismatch between the train and the test sets because the hold-out test set uses uncorrected crowdsourced labels. A model trained with a smaller α can, therefore, struggle to generalise on the hold-out test set.

Conversely, a larger α is a more conservative correction (replacing labels only at larger differences). Theoretically, a model trained with a larger α should generalise better on the test set because of a comparatively small distribution shift. This way, the model avoids training on crowdsourced labels that have a large deviation from LLM labels and, at the same time, enjoys the benefit of within-distribution crowdsourced labels that have slight deviations.

The revised dataset $\mathcal{D}'_{\text{train}} = \{(x_i, y'_i)\}$ is then used to train a PLM $\mathcal{F}_{y'}$. We hypothesise that the performance of $\mathcal{F}_{y'}$, trained on the mixture of $\tilde{y}$ and y^*, is better than $\mathcal{F}_{\tilde{y}}$, which is trained solely on $\tilde{y}$.

3) Application 2: Additional Data Labelled by LLM: Since deep learning models benefit from additional data, we propose to utilise LLM to get additional training data. While common LLM-based data augmentation techniques, such as paraphrasing [27], [45] and summarising [52], are well-documented in the literature, our approach goes a step further. Specifically, we use an LLM to label new essays following the same annotation protocol as our target domain. This method, therefore, enables the integration of any similar data points into the training process.

Mathematically, we prompt LLM to annotate new text samples u and make a new dataset $\mathcal{D}_{\text{llm}} = \{(u_i, v_i^*)\}_{i=1}^{M}$ with empathy scores v^*. These new data points could be any text similar to the essays x, but it may not have any prior empathy labels. We then annotate it in the same scale of y using LLM. This additional data is combined with $\mathcal{D}_{\text{train}}$ to create an extended training set:

$$\mathcal{D}_{\text{train+}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{llm}} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N_{\text{train}}} \cup \{(u_i, v_i^*)\}_{i=1}^{M}. \quad (2)$$

Models trained on $\mathcal{D}_{\text{train+}}$ are expected to outperform models trained solely on $\mathcal{D}_{\text{train}}$ when evaluated on the hold-out test set $\mathcal{D}_{\text{test}}$. The extended dataset would enable the model to see a more diverse set of training examples, which should improve the model's ability to generalise on unseen test data.

Similar to labelling training data required for our proposed applications, LLMs can be prompted directly to generate empathy labels for the test set D_{test}. This zero-shot prediction leverages LLM's pre-trained knowledge without requiring further fine-tuning. Compared to the other two applications, zero-shot prediction relies heavily on the inherent capabilities of the LLM.

C. Prediction using Pre-trained Language Model

Fine-tuning a pre-trained language model (PLM) is a widely adopted approach in the empathy computing literature [12]. Accordingly, we utilise the dataset refined through LLM-based approaches to fine-tune a PLM. Each text sequence x_i is first encoded into a contextual representation that serves as an aggregate sequence representation:

$$h_i^{[\text{CLS}]} = \text{PLM}(x_i; \theta), \quad (3)$$

where θ are the parameters of the PLM, and $h_i^{[\text{CLS}]} \in \mathbb{R}^d$ is the [CLS] token representation. This pooled [CLS] representation is then passed through a linear regression head to predict the continuous empathy score:

$$\hat{y}_i = \mathcal{F}(h_i^{[\text{CLS}]}; \phi) = W h_i^{[\text{CLS}]} + b, \quad (4)$$

where $\phi = W \in \mathbb{R}^{1\times d}, b \in \mathbb{R}$ denotes the learnable parameters of the linear layer. The model is trained to minimise the discrepancy between predicted scores $\hat{y}_i$ and target scores y_i^{true} across the dataset:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \ell(\hat{y}_i, y_i^{\text{true}}), \quad (5)$$

where $\ell(\cdot,\cdot)$ is the mean squared error (MSE) loss function, and N is the number of training examples. The ground truth y_i^{true} refers to the mixed labels y' for Application 1, while for Application 2, it refers to crowdsourced labels y for existing dataset $\mathcal{D}_{\text{train}}$ or LLM-provided labels v^* for additional data $\mathcal{D}_{\text{llm}}$. The evaluation is always conducted on the original held-out dataset $\mathcal{D}_{\text{test}}$.

Algorithm 1 presents the overall workflow of our proposed approaches in empathy detection. After partitioning the dataset into training, validation and test subsets, one can choose between Application 1 and 2, as they are mutually exclusive. For Application 1 (label noise correction), the LLM is queried with scale-aware prompts to generate refined labels. If the difference between the original label and the LLM-generated label exceeds a threshold, the label is updated; otherwise, the original label is retained. The revised dataset is then used for PLM fine-tuning. For Application 2 (leveraging additional unlabelled data), the LLM is queried to generate labels for the unlabelled data, which is then combined with the training set to form an extended dataset. In both cases, a PLM is fine-tuned on the revised or extended dataset. The fine-tuning involves optimising the PLM to predict empathy scores based on input text embeddings, followed by evaluating its performance on the hold-out crowdsourced test set.

IV. Experiments and Results

A. Dataset and Associated Challenges

Buechel *et al.* [57] marked an important step in understanding how individuals empathise with others or nature. They designed a crowdsourced approach where participants read newspaper articles depicting scenarios of harm to people or nature, and wrote about their emotional responses. The overarching aim was to capture individuals' reactions to adverse situations faced by others. This dataset, released in 2018,

Algorithm 1 Leveraging LLM in Empathy Detection

Require: Dataset $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$, annotation selection threshold α, additional unlabelled data $\mathcal{U} = \{u_i\}_{i=1}^{M}$
Ensure: Empathy predictions $\hat{y}$

Partition $\mathcal{D}$ into $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$ and $\mathcal{D}_{\text{test}}$
if Application 1 **then**
  go to 6
else if Application 2 **then**
  go to 12
▷ *Application 1: label noise correction* ◁
for each i in $\mathcal{D}_{\text{train}}$ **do**
  Query LLM to generate label y_i^* using scale-aware prompt
  Update label: $y_i' \leftarrow \begin{cases} y_i^*, & \text{if } |\tilde{y}_i - y_i^*| > \alpha \\ \tilde{y}_i, & \text{otherwise} \end{cases}$
Form revised dataset $\mathcal{D}'_{\text{train}} = \{(x_i, y_i')\}_{i=1}^{N_{\text{train}}}$
go to 18
▷ *Application 2: additional data labelled by LLM* ◁
for each u_i in $\mathcal{U}$ **do**
  Query LLM to generate label v_i^* for u_i using the same prompt
Form additional dataset $\mathcal{D}_{\text{llm}} = \{(u_i, v_i^*)\}_{i=1}^{M}$
Combine datasets: $\mathcal{D}_{\text{train+}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{llm}}$
go to 18
▷ *Prediction using pre-trained language model (PLM)* ◁
Fine-tune PLM $\mathcal{F}_\theta$ on $\mathcal{D}'_{\text{train}}$ or $\mathcal{D}_{\text{train+}}$:

$$\hat{y}_i = \mathcal{F}_\theta(x_i) = W h_i^{[\text{CLS}]} + b$$

Evaluate $\mathcal{F}_\theta$ on $\mathcal{D}_{\text{test}}$
return final predictions $\hat{y}$

was the first of its kind, following which subsequent datasets were built. We refer to these datasets collectively as *NewsEmp* series, as their central objective is to measure empathy elicited by newspaper articles.

The second NewsEmp dataset was released in 2022, in which Tafreshi *et al.* [19] employed 564 subjects reading 418 news articles, which led to a total of 2,655 samples distributed into training, validation and test splits. Another significant change appeared in the NewsEmp$_{23}$ dataset [58], which uses only the top 100 most negative articles from the pool of 418 news articles. Collectively, these datasets have become the most widely used dataset for benchmarking empathy detection approaches [12]. This popularity comes from their usage in the long-standing empathy detection challenge organised under the "Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA)" series [19], [37], [43], [59]. In particular, the WASSA 2021 [19] and WASSA 2022 [59] challenges utilised the NewsEmp$_{22}$ dataset, while WASSA 2023 [43] and the WASSA 2024 [37] utilised the NewsEmp$_{23}$ and the latest NewsEmp$_{24}$ datasets, respectively. Table II presents the statistics of the three datasets used in this study.

Perhaps due to the iterative nature of the datasets, there is overlap among some of these datasets. We found that the entire NewsEmp18 dataset is included in the training set of NewsEmp$_{22}$, and the entire NewsEmp$_{23}$ training and validation sets appear in the NewsEmp$_{24}$ training set. In this study, we primarily use NewsEmp$_{22}$ and NewsEmp$_{24}$ datasets and partly NewsEmp$_{23}$ datasets.

TABLE II
STATISTICS OF THE DATASETS USED IN THIS STUDY.

Name	# Train	# Validation	# Test	# Total
NewsEmp$_{22}$ [19]	1,860	270	525	2,655
NewsEmp$_{23}$ [58]	792	208	100	1,100
NewsEmp$_{24}$ [37]	1,000	63	83	1,146

Although excluding NewsEmp$_{23}$ would have been feasible, prior research [37] achieved state-of-the-art results on the NewsEmp$_{24}$ dataset by combining NewsEmp$_{22}$, NewsEmp$_{23}$ and NewsEmp$_{24}$ datasets to train their model. To ensure a fair comparison, we also report findings based on models trained using the combined three datasets.

While this combination may seem unusual due to the overlap, it can be beneficial for improving predictions on the NewsEmp$_{24}$ test split. Very likely, this test split has a similar distribution to its own training set compared to another dataset's (NewsEmp$_{22}$) training set, so including duplicated samples from the NewsEmp$_{24}$ training set allows the model to see more samples with a similar distribution.

It is worth noting that if we aim to evaluate a model on the NewsEmp$_{23}$ dataset, caution is necessary when combining datasets. One interesting finding on NewsEmp datasets is that – although not explicitly stated in the studies [37], [43], [58] reporting the datasets – 44 out of 100 test samples in the NewsEmp$_{23}$ dataset are also present in the NewsEmp$_{24}$ validation set. Due to this data leakage, a model trained on the NewsEmp$_{24}$ validation set would, therefore, inflate performance on the NewsEmp$_{23}$ test split. To verify this, we trained a model using NewsEmp$_{24}$ training and validation sets, which gives a PCC of 0.576, outperforming the state-of-the-art PCC of 0.563 in NewsEmp$_{23}$ test split [27]. To prevent misleading results in future research, we highlight this overlap here and recommend exercising caution when combining datasets. Throughout our experiments, we ensure that there is no data leakage between training/validation and test splits.

We compare the performance of our proposed LLM-based approaches across various dataset combinations, including NewsEmp$_{24}$, NewsEmp$_{23}$ and NewsEmp$_{22}$. We then benchmark our work against the evaluation metrics reported by others on the NewsEmp$_{24}$ dataset. This dataset was chosen because it is the most recent in this series, and it includes the NewsEmp$_{23}$ dataset within it. Additionally, the ground truth for the NewsEmp$_{24}$ test split is publicly available, which is essential for calculating different metrics, while the ground truth for the other datasets is publicly unavailable.

B. Evaluation Metric

Pearson correlation coefficient (PCC) is the single metric used in the literature for evaluating empathy computing models across the NewsEmp datasets [12]. While it measures linear

relationship between predicted and true values, it does not account for the *magnitude* of errors, meaning predictions can have a perfect correlation with true values while being consistently offset (*e.g.*, predictions of 1, 2, and 3 corresponding to ground truths of 5, 6, and 7 yields a PCC of 1). This issue undermines its reliability for assessing model accuracy.

While PCC has been the only metric used in empathy computing literature on NewsEmp datasets, studies on other datasets sometimes use different metrics. For example, Barros *et al.* [60], detecting empathy in an audiovisual dataset, adopted the concordance correlation coefficient (CCC) as their primary metric. Previously mentioned shortcomings of PCC could be solved using CCC, as it calculates both the linear relationship and the magnitude of prediction errors. It ensures that predictions are not only aligned with the trend of true values but also close in magnitude, penalising large errors.

Root mean square error (RMSE) appears to be another choice of evaluation as it directly captures prediction error. Overall, PCC, CCC and RMSE measure three distinct qualities of performance: PCC measures linear relationship, RMSE measures the magnitude of errors, and CCC considers both linearity and error magnitude.

C. Implementation Details

We access Llama 3 (version: `llama3-70b-8192`) through Groq API and GPT-4 (version: `gpt-4o`) through OpenAI API (last accessed on 30 December 2024). The *temperature* and *top_p* parameters of the APIs control the randomness of LLM outputs during token sampling. To ensure deterministic and consistent labels from LLMs, we set *temperature* to 0 and *top_p* to 0.01.

As pre-trained language models (PLMs), we fine-tune RoBERTa (version: `roberta-base`) [8], which has 125.7M trainable parameters, and DeBERTa (version: `deberta-v3-base`) [61], which has 184M trainable parameters, from Huggingface [62]. We apply a delayed-start early-stopping strategy that starts monitoring validation CCC after five epochs and stops training if the score does not improve for two successive epochs. Early stopping based on PCC was ineffective in our experience due to higher fluctuations of the PCC score, whereas CCC performed better due to its smoother behaviour. Since most experiments converged within 20 epochs with early stopping, we set the maximum number of training epochs to 20. Deterministic behaviour was enforced using the PyTorch Lightning framework to ensure reproducibility. We save the model checkpoint corresponding to the last epoch of training.

We adopt reported hyperparameters from the original work [8] reporting RoBERTa. Following their reported approach to fine-tuning RoBERTa for downstream tasks, we only tuned the learning rate and batch size for our task. Values of the hyperparameters are reported in Table III. While experimenting with the DeBERTa PLM, we use the same hyperparameters as used for RoBERTa.

Following [8], we report median statistics over five different random initialisations (seeds: 0, 42, 100, 999 and 1234). Since prior works on empathy computing on these datasets reported a single peak score of their model, we also report the peak score from these five runs[4]. All experiments are conducted in Python 3 running on a single AMD Instinct™ MI250X GPU (64 GB).

TABLE III
HYPERPARAMETERS FOR MODEL TRAINING.

Pramerter	Value	Pramerter	Value
Optimiser	AdamW	Learning rate scheduler	Linear
Learning rate	3e-5	Warmup ratio	0.06
AdamW (β_1, β_2)	(0.9, 0.98)	Batch size	16
AdamW ϵ	1e-6	Maximum epochs	20
Weight decay	0.1	Max sequence length	512

D. Main Results

This section presents the quantitative results of our proposed applications of LLM as a service (LLMaaS) in empathy prediction. Unless otherwise stated, results are reported primarily using the RoBERTa PLM, with DeBERTa results explicitly specified.

1) Noise Mitigation: We first show evidence of noise in the NewsEmp_{24} dataset. Table IV illustrates a comparison between human participants' and LLMs' assessments of empathy on a scale of 1 (lowest empathy) to 7 (highest empathy) in two example essays. It demonstrates interesting disparities between crowdsourced and LLM evaluations – for instance, in one essay expressing deep emotional concern for affected people and children, the human rater assigned a relatively low score of 1.0, while the Llama and GPT LLMs rated it much higher at 6.4 and 6.08, respectively. Conversely, a less empathic account of a mining disaster received the maximum possible empathy score (7.0) from human raters but a much lower 1.83 and 1.67 from the Llama and GPT LLMs, respectively. Interestingly, both LLMs, despite differences in size and provider, produce highly consistent annotations, which further underscores the potential inaccuracies of the crowdsourced annotation.

We evaluate our proposed LLMaaS application for noise mitigation in three dataset configurations: NewsEmp_{24} alone, NewsEmp_{24+22}, and $\text{NewsEmp}_{24+23+22}$. While combining the datasets, we combine their training and validation splits of the additional dataset with the training split of the base dataset for training the model. For example, the NewsEmp_{24+22} experimental setup uses the training split of NewsEmp_{24} and training and validation splits of the NewsEmp_{22} dataset to train the model. In all cases, the model training is optimised for the validation split of the NewsEmp_{24} dataset and finally evaluated on the hold-out NewsEmp_{24} test split.

As presented in Table V, our LLM-based noise mitigation approach demonstrates consistent performance improvements across all configurations. For the NewsEmp_{24} dataset configuration, we report results with both Llama- and GPT-generated labels. While we choose RoBERTa as the primary

[4]We define *peak* score as the best score (maximum PCC, maximum CCC or minimum RMSE) across five random runs within a *single* experimental setup. Another related terminology used throughout this paper is the *best* score, which refers to the best scores across *different* experimental setups.

TABLE IV
EXAMPLES OF MISLABELLED CROWDSOURCED ANNOTATION, DEVIATING FROM BATSON'S DEFINITION OF EMPATHY. THE FIRST EXAMPLE SHOWS AN ESSAY WITH EMPATHIC ELEMENTS, BUT THE PARTICIPANT'S ANNOTATION INDICATES THE LOWEST EMPATHY. THE SECOND EXAMPLE HAS THE HIGHEST EMPATHY SCORE DESPITE THE ESSAY LACKING EMPATHIC CONTENT. LLM LABELS APPEAR ACCURATE AND CONSISTENT BETWEEN LLAMA AND GPT.

Essay	Crowd	Llama	GPT
"After reading the article, my heart just breaks for the people that are affected by this. Not only are innocent people being killed daily but also little children as well as babies. These children do not deserve this and it's sad because they have their whole lives ahead of them. I really hope that war will end one day although it is looking unlikely."	1.0	6.4	6.08
"I read the article on the China mining disaster. There were 33 miners trapped in the mine. Only two of them survived. Officials stated whoever was responsible would be punished. Smaller mines were shut down immediately until further notice. China has always been known for the deadliest mining."	7.0	1.83	1.67

Empathic expressions are highlighted in blue.
Labels are in a continuous range from 1 to 7, where 1 and 7 refer to the lowest and highest empathy, respectively.

TABLE V
RESULTS OF OUR LLM-BASED NOISE MITIGATION APPROACH, EVALUATED ON THE NEWSEMP$_{24}$ TEST SET.

Labels	α	PCC ↑	CCC ↑	RMSE ↓
Data: NewsEmp$_{24}$, PLM: RoBERTa				
CS	–	0.331(0.378)	0.307(0.329)	1.656(**0.066**)
CS & Llama	3.5	0.453(0.462)	**0.435**(0.455)	1.604(0.087)
	4.0	0.384(0.464)	0.378(0.454)	1.647(0.075)
	4.5	0.421(0.482)	0.392(0.463)	1.566(0.098)
CS & GPT	3.5	**0.473**(0.509)	0.431(**0.496**)	**1.558**(1.499)
	4.0	0.415(**0.519**)	0.398(0.482)	1.601(1.461)
	4.5	0.370(0.422)	0.325(0.400)	1.646(1.532)
Data: NewsEmp$_{24}$, PLM: DeBERTa				
CS	–	0.447(0.481)	0.398(0.420)	1.481(1.441)
CS & Llama	3.5	0.502(0.516)	0.450(0.511)	1.531(1.445)
	4.0	0.536(0.576)	0.500(0.518)	1.413(1.369)
	4.5	0.476(0.525)	0.433(0.479)	1.470(1.398)
CS & GPT	3.5	0.529(0.568)	0.489(0.536)	1.419(1.393)
	4.0	**0.558**(**0.596**)	**0.533**(**0.571**)	**1.408**(**1.326**)
	4.5	0.526(0.554)	0.484(0.521)	1.425(1.341)
Data: NewsEmp$_{24+22}$, PLM: RoBERTa				
CS	–	0.536(0.597)	0.461(0.505)	1.356(**0.042**)
CS & Llama	0.0	0.483(0.504)	0.445(0.480)	1.655(1.628)
	0.5	0.488(0.503)	0.445(0.481)	1.646(1.617)
	1.0	0.479(0.491)	0.432(0.451)	1.756(1.664)
	1.5	0.474(0.498)	0.434(0.448)	1.694(1.583)
	2.0	0.455(0.519)	0.421(0.453)	1.661(1.575)
	2.5	0.502(0.547)	0.447(0.456)	1.622(1.585)
	3.0	0.543(0.571)	0.507(0.512)	1.461(1.435)
	3.5	0.558(0.612)	0.496(0.559)	1.389(0.084)
	4.0	**0.589**(**0.627**)	**0.516**(0.563)	**1.338**(0.054)
	4.5	0.551(0.620)	0.478(**0.575**)	1.378(0.100)
	5.0	0.551(0.605)	0.478(0.520)	1.348(1.265)
	5.5	0.542(0.604)	0.464(0.516)	1.352(1.265)
	6.0	0.536(0.597)	0.461(0.505)	1.356(1.274)
Data: NewsEmp$_{24+23+22}$, PLM: RoBERTa				
CS	–	0.528(0.551)	0.469(0.498)	1.380(0.086)
CS & Llama	3.5	0.556(0.573)	0.511(0.552)	1.381(0.029)
	4.0	**0.574**(0.582)	**0.529**(0.548)	**1.333**(**0.021**)
	4.5	0.548(**0.648**)	0.479(**0.597**)	1.346(0.092)

CS – crowdsourced labels; α – annotation selection threshold.
Reported metrics are in `median(peak)` format, calculated from five random initialisations.
Boldface and underline texts indicate the best and the second-best scores, respectively.

PLM, we also report results with DeBERTa PLM in this setup, which shows even better performance improvement. Specifically, GPT labels at $\alpha = 4.0$ achieve the best median PCC (0.558), CCC (0.533) and RMSE (1.408) scores. Overall, the performance improvement between Llama and GPT labels is comparable, with each achieving the best results in certain metrics. Since Llama is an open-weight LLM and freely available, we proceed with Llama for the remaining experiments in this application scenario.

Including the NewsEmp$_{22}$ dataset enhances performance further, with $\alpha = 4.0$ yielding the best median PCC (0.589) and median CCC (0.516). When combined with NewsEmp$_{23}$, the baseline metrics remain comparable, but $\alpha = 4.0$ again delivers the highest median PCC (0.574) and median CCC (0.529), with RMSE achieving its lowest value of 1.333. Considering peak scores instead of median statistics across five runs, our approach also outperforms the baseline model by achieving the peak PCC of 0.648, CCC of 0.597 and RMSE of 0.021. As illustrated earlier in Fig. 1, the performance improvements are *statistically* significant.

The value of the threshold α controls the proportion of LLM and crowdsourced labels. As demonstrated earlier, a smaller value of α means having a higher amount of LLM labels, which may hurt the model's performance in the test set. This hypothesis is verified in Table V's results on varying α on the NewsEmp$_{24+22}$ scenario, which shows that $\alpha = 3.5 \sim 4.5$ provides the best performance across the three dataset configurations.

Our noise mitigation approach outperforms the baseline in terms of median PCCs, CCCs and RMSEs, as well as peak PCCs and CCCs across all four configurations. Out of 24 test cases[5], only two cases (RoBERTa on NewsEmp$_{24}$ and NewsEmp$_{24+22}$), show a better peak RMSE achieved by the baseline. This discrepancy can be primarily attributed to how training was controlled: we applied early stopping based on CCC to mitigate overfitting. Since CCC and RMSE are not strongly correlated [63], [64], early stopping by CCC does not necessarily optimise RMSE.

Recent works in affective computing, including emotion recognition [65] and empathy computing [12], preferred correlation-based metrics over error-based metrics. Our findings align with this trend: although the baseline approach incidentally achieves a better peak RMSE in two isolated runs, our method demonstrates more consistent and robust performance across all metrics, including median RMSE across five runs, which reflects typical behaviour rather than outliers. In

[5]2 types (median, peak) × 3 metrics × 4 configurations in Table V.

TABLE VI
EFFECT OF ADDITIONAL LABELLED DATA.

Training data	PCC ↑	CCC ↑	RMSE ↓
PLM: RoBERTa			
NewsEmp_{24}	0.331(0.378)	0.307(0.329)	1.656(0.066)
+ Crowd-labelled NewsEmp_{22}	0.485(**0.594**)	0.439(0.480)	**1.417**(0.093)
+ Llama-labelled NewsEmp_{22}	**0.513**(0.571)	**0.490**(**0.523**)	1.484(**0.059**)
+ GPT-labelled NewsEmp_{22}	0.495(0.549)	0.446(0.455)	1.581(1.514)
PLM: DeBERTa			
NewsEmp_{24}	0.447(0.481)	0.398(0.420)	1.481(1.441)
+ Crowd-labelled NewsEmp_{22}	0.564(**0.638**)	0.478(0.566)	**1.329**(**1.233**)
+ Llama-labelled NewsEmp_{22}	**0.581**(0.601)	**0.535**(**0.584**)	1.445(1.366)
+ GPT-labelled NewsEmp_{22}	0.554(0.596)	0.493(0.534)	1.468(1.391)
PLM: RoBERTa			
NewsEmp_{22}	0.459(0.477)	0.363(0.411)	1.776(0.046)
+ Crowd-labelled NewsEmp_{24}	0.467(0.478)	0.392(**0.435**)	1.756(0.062)
+ Llama-labelled NewsEmp_{24}	**0.496**(**0.519**)	**0.429**(0.434)	**1.729**(**0.034**)

All evaluations use test splits, except for NewsEm22, where CCC and RMSE are computed on the validation split due to unavailable test labels.
Reported metrics are in `median(peak)` format, calculated across five random initialisations.
Boldface and underline texts indicate the best and the second-best scores, respectively.

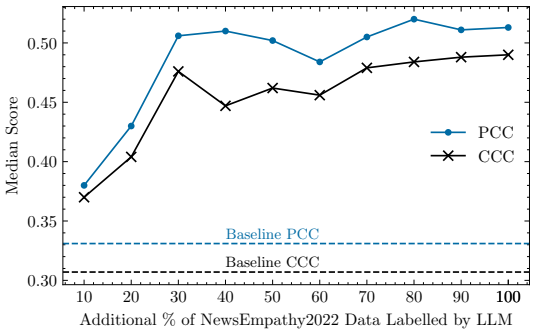

Fig. 3. Median performance in the NewsEmp_{24} test set with gradual increase of additional LLM-labelled data. *Baseline* scores refer to the scores achieved using only NewsEmp_{24} data.

terms of the choice of the primary metric, PCC only captures linear correlation but not error magnitude, while RMSE only captures error magnitude but not correlation; therefore, our recommendation is to consider CCC as the primary metric, as it captures the best of both worlds – both linear correlation and error magnitude.

2) Additional Data Labelled by LLM: The first application described above demonstrates that additional training data helps achieve better performance. However, we may not always have the flexibility of having extra data labelled by human participants. This application, therefore, explores whether additional data labelled by LLM could help.

We evaluate this application in two configurations: either NewsEmp_{24} or NewsEmp_{22} as the base dataset. While evaluating the model on the NewsEmp_{24} dataset, we consider the NewsEmp_{22} dataset as additional data and vice versa. Since we have both crowdsourced and LLM-generated labels for the additional data, we compare whether the additional data is labelled by (1) human participants or (2) LLM.

Table VI reports the performance in both settings. When trained a RoBERTa model with the NewsEmp_{24} dataset alone, it achieved a median PCC of 0.331 and 0.307 CCC. Additional NewsEmp_{22} dataset labelled by human participants boosted performance to 0.485 PCC, 0.439 CCC and 1.417 RMSE. The same NewsEmp_{22} dataset – labelled by Llama LLM – makes the highest median PCC of 0.513 and CCC of 0.490 while maintaining a competitive RMSE of 1.484. For this NewsEmp_{24} setup, we further report results with the DeBERTa PLM, which shows even better performance across different metrics. Like our earlier experiment (Application 1), the use of either Llama or GPT LLMs yields similar performance in this setup, and we proceed with Llama LLMs for the rest of the experiments.

Experiments using NewsEmp_{22} as the base dataset exhibited a similar trend to those with NewsEmp_{24} as the base dataset. The model trained with LLM-generated labels achieves the best PCC (0.496) and CCC (0.429), as well as the lowest RMSE (1.729). Notably, the human-labelled data shows improvement over the base dataset but falls short of the performance achieved with LLM-labelled data. Overall, LLM labels are as good as crowdsourced labels, and additional data, either crowdsourced or LLM-labelled, boosts the performance.

Compared to our first application scenario (mixed labels to reduce label noise), performance improvement from baseline in this application is statistically more significant across all three evaluation metrics (Fig. 1). In particular, this application scenario demonstrates the highest level of statistically significant improvements in terms of PCC and CCC. This is likely because, in this scenario, the labels of the base training data remain unchanged, which is presumably of a similar distribution to the hold-out test set. The additional data provides extra supervision, which helps achieve a better score. Results using additional data labelled by LLM are better than using human labelling in most cases (12 out of 18 test cases), likely because of the higher quality of labels from LLM.

To understand how the amount of additional data affects the performance, we gradually increase the amount of additional data and report model performance (Fig. 3). In each case, we randomly sample a percentage of the additional data ranging from 10% to 100%. Surprisingly, the performance increases most rapidly from 10 to 30%, after which the improvement slows down.

Fig. 4 presents 3D t-SNE visualisations of embeddings derived from different labelling schemes, alongside their clustering performance measured by the mean Silhouette score [66] on the embeddings from PLM. Given the continuous nature of empathy labels in the datasets, we discretise them into six bins (1–2, 2–3, ..., 6–7) to calculate the metric. This score ranges from -1 to $+1$, with -1 being *"misclassified"* and $+1$ being *"well-clustered"* [66].

Embeddings from crowdsourced labels exhibit a slightly dispersed distribution visually and the lowest Silhouette score (-0.005), which suggests many samples are mislabelled (*i.e.*, label noise). In contrast, embeddings based on LLM labels display smoother distributions (especially the rightmost

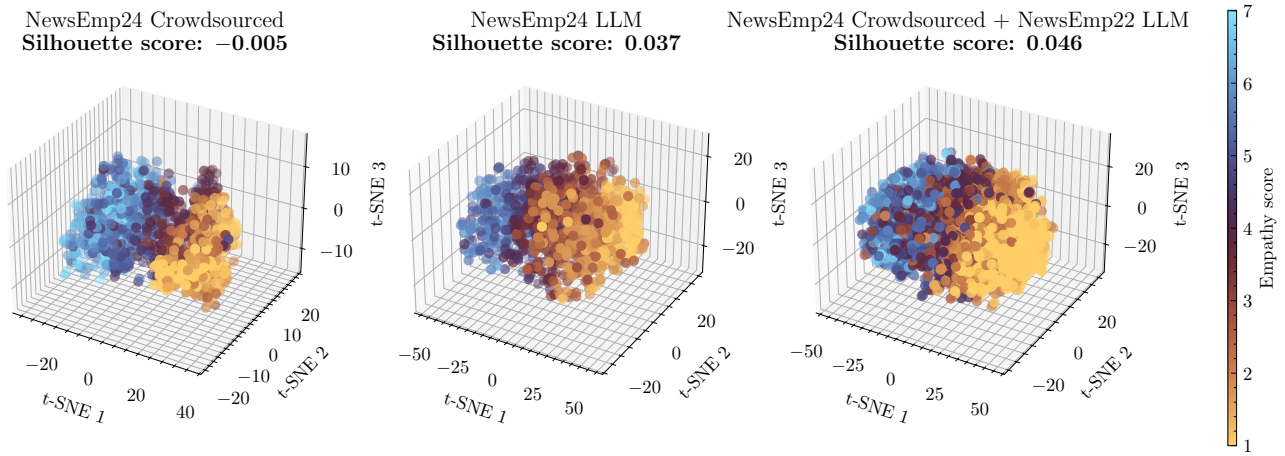

Fig. 4. 3D t-SNE visualisation and Silhouette scores on CLS embeddings from pre-trained language models fine-tuned using crowdsourced labels (**left**), LLM-generated labels (**middle**) and crowdsourced + additional LLM-labelled data (**right**). Continuous empathy labels are discretised into six bins to calculate Silhouette scores (higher is better), which suggests better clustering after integrating LLM-generated labels.

TABLE VII
ZERO-SHOT EMPATHY PREDICTION USING LLMS.

Dataset	LLM	Split	PCC ↑	CCC ↑	RMSE ↓
NewsEmp_{24}	Llama	Test	0.441	0.436	1.731
		Validation	0.502	0.457	1.952
	$\text{Llama}_{\text{Plain}}$	Test	0.405	0.358	1.859
	GPT	Test	0.581	0.489	1.715
		Validation	0.480	0.375	2.038
NewsEmp_{23}	Llama	Test	0.380	–	–
		Validation	0.108	0.108	2.18
NewsEmp_{22}	Llama	Test	0.517	–	–
		Validation	0.579	0.573	1.728

$\text{Llama}_{\text{Plain}}$: Llama with the plain prompt. All other entries use the scale-aware prompt.
Ground truth of the NewsEm22 and NewsEmp_{23} test splits are unavailable to calculate CCC and RMSE.

plot) and higher Silhouette scores (0.037 and 0.046)[6], which suggests many of the mislabelled samples are now correctly labelled (*i.e.*, reduction of label noise). The smoothest distribution and the highest Silhouette score on the NewsEmp_{24} Crowdsourced + NewsEmp_{22} LLM configuration can be attributed to extra supervision from the LLM-generated labels.

E. Zero-Shot Prediction & Demographic Biases

The most direct application of LLM is to predict empathy in a zero-shot manner, *i.e.*, without any training or fine-tuning of the LLM. Table VII reports the performances of zero-shot prediction across all validation and test splits.

On the NewsEmp_{24} test split, we compare plain prompting with our proposed scale-aware prompting scheme. The improved performance with the scale-aware prompt (PCC: 0.405 → 0.441; CCC: 0.358 → 0.436; RMSE: 1.859 → 1.731) demonstrates its effectiveness. Accordingly, we adopt the scale-aware prompt as the default in all our experiments.

We further examine how the agreement between crowdsourced annotation and LLM annotation varies across different demographic groups. For this analysis, we combine training, validation and test splits of the NewsEmp_{24} dataset and compare crowdsourced and Llama-generated labels. As demonstrated in Fig. 5, both have similar levels of CCC in gender and education demographics. However, CCC changed wildly across race, age and income groups. In particular, it went to negatives in two race groups: "Hispanic/Latino" and "Other" categories, noting that there are only four samples in the "Other" category of race.

In terms of the number of samples across different demographics, we see that certain demographic groups (*e.g.*, "4-year bachelor's degree" education, "White" race, and "31-40" age groups) are highly represented compared to their counterparts. Some demographics, for example, "Less than high school" education and "Native American / American Indian" are not represented in the dataset at all. Such imbalances can presumably introduce bias in the empathy detection model built from such biased datasets.

LLMs designed for empathy detection may exhibit biases across different demographic groups [17]. A proper assessment of such bias would require accurate and unbiased ground truth labels to compare with. Note that Fig. 5 compares LLM outputs against potentially noisy crowdsourced labels and so may not offer an objective assessment of bias. Instead, it reflects potential *sources* of bias that may influence an empathy detection model, insofar as biased training data can propagate bias into the model's predictions.

Our prompt to the LLM is *demographic-unaware*, meaning it does not include any explicit demographic information. Theoretically, LLM outputs through such an unaware prompt should be less biased (because it is a blind evaluation) compared to a demographic-aware prompt [17]. Gabriel *et al.* [17] argued that a demographic-aware prompt may also result in less biased assessment, since the LLMs are being alerted

[6]Note that although Fig. 4 shows relative differences in Silhouette scores across embeddings, their absolute values remain low, presumably due to the discretisation of continuous labels (the clustered labels are not well-separated groups).

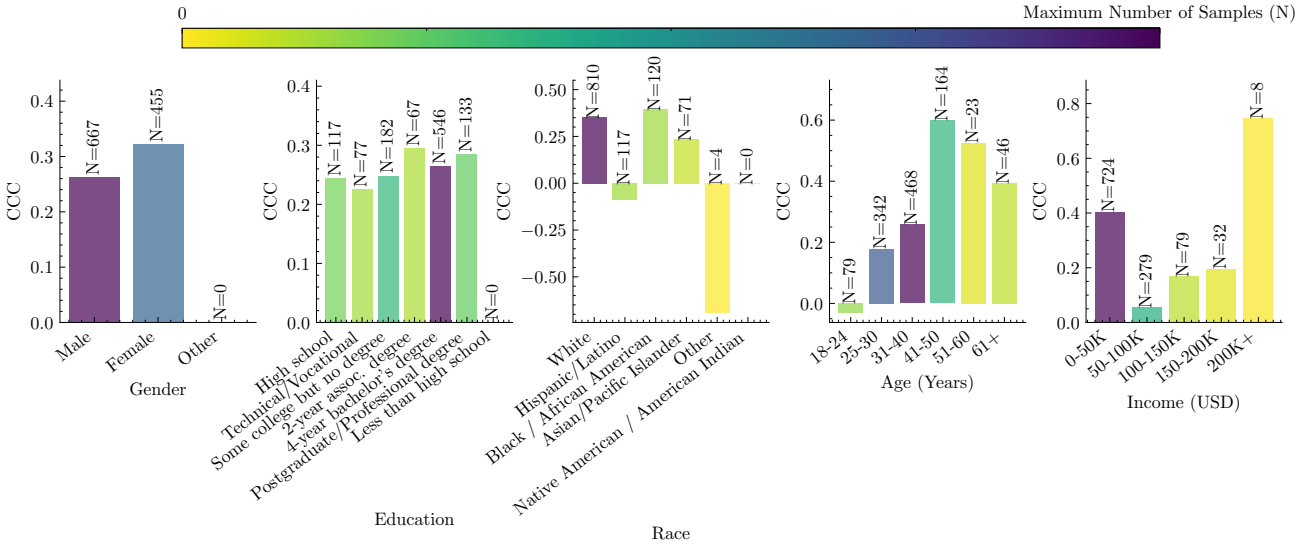

Fig. 5. Number of samples and zero-shot (Llama) prediction performance across different demographic groups in the NewsEmp$_{24}$ dataset. CCC varies rapidly across different racial groups.

TABLE VIII
COMPARISON OF OUR PROPOSED MODEL WITH THE LITERATURE ON THE NEWSEMP$_{24}$ TEST DATASET.

Approach	Base Model (Ref.)	PCC ↑	CCC ↑	RMSE ↓
Training	BERT [53]	0.290	–	–
	MLP [67]	0.345	–	–
	RoBERTa [49]	0.375	–	–
	Not mentioned [68]	0.390	–	–
	Llama 3 8B [2]	0.474	–	–
	RoBERTa [37]	0.629	–	–
	RoBERTa [37][a]	0.607	0.498	**0.075**
	RoBERTa (**Ours**)	**0.648**	**0.597**	0.092
Zero-shot	GPT 3.5 [3]	0.523	–	–
	GPT 4 (**Ours**)	0.581	0.489	1.715

[a] Our implementation of the earlier SOTA work [37].

TABLE IX
CONSISTENCY AND INTER-RATER RELIABILITY AMONG LLAMA, GPT AND CROWDSOURCED ANNOTATIONS ON THE NEWSEMP$_{24}$ TRAINING SET. THE LOW RELIABILITY BETWEEN LLMS AND CROWDSOURCED ANNOTATIONS, CONTRASTED WITH THE HIGH RELIABILITY BETWEEN TWO DIFFERENT LLMS, MAY SUGGEST THAT THE CROWDSOURCED ANNOTATIONS ARE NOISY.

Annotator 1	Annotator 2	Krippendorff's Alpha	MAE $\pm$ SD
Llama	Llama	0.99	0.10 ± 0.21
Llama	GPT	0.80	0.78 ± 0.70
Llama	Crowd	0.27	1.72 ± 1.34
GPT	Crowd	0.19	1.81 ± 1.27

MAE – mean absolute error; SD – standard deviation of absolute error

to potential bias. They found mixed outcomes on mitigating bias through demographic-aware and unaware prompts across different LLMs [17]. Mitigating bias is critical for real-world deployment, and so future exploration towards a universally applicable prompting strategy is warranted to mitigate demographic biases.

F. Comparison with the Literature

A quantitative comparison between our proposed framework and other empathy detection works in the literature on the NewsEmp$_{24}$ dataset is presented on Table VIII. The best performance in the literature is 0.629 PCC [37], while our best performance is 0.648 PCC. Both Giorgi *et al.* [37] and we use the same amount of training data – combined NewsEmp$_{24+23+22}$.

The reported results in the literature are in terms of a single evaluation metric, which suggests the peak performance of their model. To compare in terms of other evaluation metrics in a similar setting of ours (five random initialisations), we implemented the state-of-the-art work [37]. The mismatch between our implementation and Giorgi *et al.* [37]'s reported result (0.607 vs 0.629) is likely due to hyperparameter choice. Having no public implementation of Giorgi *et al.* [37], we chose default hyperparameters, apart from the minimal amount of hyperparameter details reported in their work. Our approach outperforms Giorgi *et al.* [37]'s results in terms of both PCC and CCC (Table VIII).

G. Consistency and Inter-Rater Reliability

LLMs are known to produce varying outputs across different API calls [69]. This variability could raise concerns about using LLM to label data as well as evaluating on the test set. We calculate two types of consistency: *intra-LLM* consistency, which evaluates whether annotations generated by the same LLM model remain consistent across multiple API calls, and *inter-LLM* consistency, which assesses whether annotations are consistent between two different LLMs.

To assess intra-LLM consistency, we label the NewsEmp$_{24}$ training set (1,000 samples) twice in the Llama LLM, using separate independent API calls. The results (Table IX) demonstrate *"almost perfect"* agreement between the two annotation rounds, with a Krippendorff's Alpha (K-Alpha) [70] score of

0.99. Between GPT and LLM annotations (inter-LLM), a K-Alpha of 0.80 is achieved, which lies on the boundary between *"substantial"* and *"almost perfect"* reliability [70]. Such a high level of consistency, including inter-LLM consistency, suggests the effectiveness of our prompting strategy, which clearly specifies the expectations from the LLM.

As presented in Table IX, the inter-rater reliability between LLMs and crowdsourced annotations is notably lower than the reliability observed between LLMs. It potentially supports our hypothesis that crowdsourced annotations are inherently noisy.

H. Human Preference Study: Crowdsourced vs LLM Labels

To evaluate the credibility of LLM-generated empathy scores relative to crowdsourced annotations, we conducted a small-scale human assessment study involving three co-authors (all PhD holders, including two mid-career and one senior academic) as independent assessors. The corresponding author designed the experiment, while the assessors were blind to the origin of each label (LLM or crowdsourced). We selected 20 samples that exhibited the largest disagreement between LLM and crowdsourced annotations to better test discernibility. Each assessor was shown an essay along with two empathy scores (randomised in order) and asked to choose the score that best reflected the empathy expressed in the essay, following Batson's definition of empathy.

Across the 20 samples, 8 were unanimously rated in favour of the LLM-generated labels by all three assessors. In 10 other samples, two out of three assessors preferred the LLM labels. Only in 2 cases did the majority (2 of 3) select the crowdsourced label. Based on majority voting, LLM-generated labels were preferred in 18 out of 20 instances, suggesting that, at least in these high-discrepancy cases, LLM annotations aligned more closely with human expert judgment than the original crowdsourced labels.

I. Limitations and Future Work

As we note the inherent biases in LLM, it is crucial to exercise caution when using LLM-generated labels if such a system is to be deployed in real life. Zero-shot predictions are likely to exhibit greater bias, as LLMs may inherit biases from their training data. Therefore, downstream models should be trained on diverse and representative datasets that reflect the demographics in which empathy would be detected.

As this paper investigates how LLM-generated labels can help measure empathy by smaller PLMs, we restrict the LLMs to generating labels only. However, incorporating the reasoning behind these labels could enhance explainability and potentially further improve label quality. Future work may explicitly prompt LLMs to demand justification, employ chain-of-thought prompting to elicit reasoning [71], or leverage LLMs having implicit reasoning capability (*e.g.*, OpenAI o3). Lastly, as all experiments are conducted on English datasets (NewsEmp series), exploring the multilingual adaptability of our methods remains an important direction for future work.

V. Conclusion

This work demonstrates the potential of large language models (LLMs) in addressing challenges in empathy computing through two *in-vitro* applications: label noise reduction and training data expansion. Both applications resulted in statistically significant performance gains over baseline methods. The proposed framework outperformed state-of-the-art methods, achieving new benchmarks on a public empathy dataset with a Pearson correlation coefficient (PCC) of 0.648, among other metrics. Beyond the empirical results, this paper contributes a critical rethinking of evaluation practices in empathy computing, advocating for the adoption of the concordance correlation coefficient (CCC). The novel scale-aware prompting technique introduced here ensures alignment between LLM annotations and theoretical annotation protocols. We further highlight biases in the dataset across different demographic groups. Similar to the empathy detection dataset addressed in this paper, many other tasks, such as detecting depression, anxiety and mental health conditions, rely on questionnaire-based self-annotations. The proposed approach, therefore, opens exciting avenues for leveraging LLMs as complementary tools to enhance model training across different domains.

Acknowledgement

This work was supported by resources provided by the Pawsey Supercomputing Research Centre with funding from the Australian Government and the Government of Western Australia.

References

[1] Z. Ma, W. Wu, Z. Zheng, *et al.*, "Leveraging speech ptm, text llm, and emotional tts for speech emotion recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 146–11 150. DOI: 10.1109/ICASSP48485.2024.10445906.

[2] T. Li, N. Rusnachenko, and H. Liang, "Chinchunmei at WASSA 2024 empathy and personality shared task: Boosting LLM's prediction with role-play augmentation and contrastive reasoning calibration," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 385–392.

[3] H. Kong and S. Moon, "RU at WASSA 2024 shared task: Task-aligned prompt for predicting empathy and distress," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 380–384.

[4] T. Sun, Y. Shao, H. Qian, X. Huang, and X. Qiu, "Black-box tuning for language-model-as-a-service," in *International Conference on Machine Learning*, PMLR, 2022, pp. 20 841–20 855.

[5] Y. Hu, Q. Chen, J. Du, *et al.*, "Improving large language models for clinical named entity recognition via prompt engineering," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1812–1820, Jan. 2024. DOI: 10.1093/jamia/ocad259.

[6] H. Fei, B. Li, Q. Liu, L. Bing, F. Li, and T.-S. Chua, "Reasoning implicit sentiment with chain-of-thought prompting," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 1171–1182. DOI: 10.18653/v1/2023.acl-short.101.

[7] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3774–3782. DOI: 10.1109/ICCV.2017.405.

[8] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized bert pretraining approach," 2019.

[9] M. Mazumder, C. Banbury, X. Yao, *et al.*, "Dataperf: Benchmarks for data-centric ai development," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, Curran Associates, Inc., 2023, pp. 5320–5347.

[10] R. S. Geiger, K. Yu, Y. Yang, *et al.*, "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*, 2020, pp. 325–336. DOI: 10.1145/3351095.3372862.

[11] M. L. Hoffman, *Empathy and Moral Development: Implications for Caring and Justice*. Cambridge University Press, 2000. DOI: 10.1017/CBO9780511805851.

[12] M. R. Hasan, M. Z. Hossain, S. Ghosh, A. Krishna, and T. Gedeon, *Empathy detection from text, audiovisual, audio or physiological signals: A systematic review of task formulations and machine learning methods*, 2024.

[13] C. D. Batson, J. Fultz, and P. A. Schoenrade, "Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences," *Journal of personality*, vol. 55, no. 1, pp. 19–39, 1987. DOI: 10.1111/j.1467-6494.1987.tb00426.x.

[14] F. Charlier, M. Weber, D. Izak, *et al.*, *Statannotations*, version v0.6, Oct. 2022. DOI: 10.5281/zenodo.7213391.

[15] B. D. Jani, D. N. Blane, and S. W. Mercer, "The role of empathy in therapy and the physician-patient relationship," *Complementary Medicine Research*, vol. 19, no. 5, pp. 252–257, 2012. DOI: 10.1159/000342998.

[16] K. Aldrup, B. Carstensen, and U. Klusmann, "Is empathy the key to effective teaching? a systematic review of its association with teacher-student interactions and student outcomes," *Educational Psychology Review*, vol. 34, no. 3, pp. 1177–1216, 2022. DOI: 10.1007/s10648-021-09649-y.

[17] S. Gabriel, I. Puri, X. Xu, M. Malgaroli, and M. Ghassemi, "Can AI relate: Testing large language model response for mental health support," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds., Miami, Florida, USA, Nov. 2024, pp. 2206–2221. DOI: 10.18653/v1/2024.findings-emnlp.120.

[18] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021. DOI: 10.1145/3446776.

[19] S. Tafreshi, O. De Clercq, V. Barriere, S. Buechel, J. Sedoc, and A. Balahur, "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*, Online: Association for Computational Linguistics, Apr. 2021, pp. 92–104.

[20] S. Mohammad and P. Turney, "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*, 2010, pp. 26–34.

[21] K. B. Sheehan, "Crowdsourcing research: Data collection with amazon's mechanical turk," *Communication Monographs*, vol. 85, no. 1, pp. 140–156, 2018. DOI: 10.1080/03637751.2017.1342043.

[22] R. Jia, Z. R. Steelman, and B. H. Reich, "Using mechanical turk data in is research: Risks, rewards, and recommendations," *Communications of the Association for Information Systems*, vol. 41, no. 1, p. 14, 2017.

[23] J. L. Huang, P. G. Curran, J. Keeney, E. M. Poposki, and R. P. DeShon, "Detecting and deterring insufficient effort responding to surveys," *Journal of Business and Psychology*, vol. 27, pp. 99–114, 2012.

[24] W. O'Brochta and S. Parikh, "Anomalous responses on amazon mechanical turk: An Indian perspective," *Research & Politics*, vol. 8, no. 2, 2021. DOI: 10.1177/20531680211016971.

[25] M. Niu, M. Jaiswal, and E. M. Provost, "From text to emotion: Unveiling the emotion annotation capabilities of llms," in *Proc. Interspeech 2024*, 2024, pp. 2650–2654. DOI: 10.21437/Interspeech.2024-2282.

[26] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Want to reduce labeling cost? GPT-3 can help," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4195–4205. DOI: 10.18653/v1/2021.findings-emnlp.354.

[27] M. R. Hasan, M. Z. Hossain, T. Gedeon, and S. Rahman, "LLM-GEm: Large language model-guided prediction of people's empathy levels towards newspaper article," in *Findings of the Association for Computational Linguistics: EACL 2024*, Y. Graham and M. Purver, Eds., St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2215–2231.

[28] A. Grattafiori, A. Dubey, A. Jauhri, *et al.*, *The Llama 3 herd of models*, 2024.

[29] OpenAI, J. Achiam, S. Adler, *et al.*, *GPT-4 technical report*, 2024.

[30] Y. Wang, T. Baldwin, and K. Verspoor, "Noisy label regularisation for textual regression," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, *et al.*, Eds., Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 4228–4240.

[31] E. Englesson and H. Azizpour, "Robust classification via regression for learning with noisy labels," in *The Twelfth International Conference on Learning Representations*, 2024.

[32] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari, "Learning with noisy labels," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26, Curran Associates, Inc., 2013.

[33] S. Garg, G. Ramakrishnan, and V. Thumbe, "Towards robustness to label noise in text classification via noise modeling," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 3024–3028. DOI: 10.1145/3459637.3482204.

[34] P. Li, X. He, X. Cheng, *et al.*, "An improved categorical cross entropy for remote sensing image classification based on noisy labels," *Expert Systems with Applications*, vol. 205, p. 117 296, 2022. DOI: 10.1016/j.eswa.2022.117296.

[35] B. Zhang, Y. Wang, W. Hou, *et al.*, "Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 18 408–18 419.

[36] K. Sohn, D. Berthelot, N. Carlini, *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 596–608.

[37] S. Giorgi, J. Sedoc, V. Barriere, and S. Tafreshi, "Findings of WASSA 2024 shared task on empathy and personality detection in interactions," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 369–379.

[38] S. Qian, C. Orašan, D. Kanojia, H. Saadany, and F. Do Carmo, "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*, 2022, pp. 271–275. DOI: 10.18653/v1/2022.wassa-1.29.

[39] A. Lahnala, C. Welch, and L. Flek, "CAISA at WASSA 2022: Adapter-tuning for empathy prediction," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2022, pp. 280–285. DOI: 10.18653/v1/2022.wassa-1.31.

[40] Y. Chen, Y. Ju, and S. Kübler, "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*, 2022. DOI: 10.18653/v1/2022.wassa-1.21.

[41] F. M. Plaza-del-Arco, J. Collado-Montañez, L. A. Ureña, and M.-T. Martín-Valdivia, "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*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 239–244. DOI: 10.18653/v1/2022.wassa-1.23.

[42] Y. Wang, J. Wang, and X. Zhang, "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*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 526–530.

[43] V. Barriere, J. Sedoc, S. Tafreshi, and S. Giorgi, "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, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 511–525.

[44] F. Gruschka, A. Lahnala, C. Welch, and L. Flek, "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*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 553–557.

[45] H. Vasava, P. Uikey, G. Wasnik, and R. Sharma, "Transformer-based architecture for empathy prediction and emotion classification," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2022, pp. 261–264. DOI: 10.18653/v1/2022.wassa-1.27.

[46] A. Kulkarni, S. Somwase, S. Rajput, and M. Marathe, "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*, Online: Association for Computational Linguistics, Apr. 2021, pp. 105–111.

[47] A. S. Srinivas, N. Barua, and S. Pal, "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*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 542–547.

[48] X. Lu, Z. Li, Y. Tong, Y. Zhao, and B. Qin, "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*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 574–580.

[49] R. Frick and M. Steinebach, "Fraunhofer SIT at WASSA 2024 empathy and personality shared task: Use of sentiment transformers and data augmentation with fuzzy labels to predict emotional reactions in conversations and essays," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 435–440.

[50] S. Ghosh, D. Maurya, A. Ekbal, and P. Bhattacharyya, "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*, 2022, pp. 255–260. DOI: 10.18653/v1/2022.wassa-1.26.

[51] Y. Butala, K. Singh, A. Kumar, and S. Shrivastava, "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*, Online: Association for Computational Linguistics, Apr. 2021, pp. 274–280.

[52] M. R. Hasan, M. Z. Hossain, T. Gedeon, S. Soon, and S. Rahman, "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*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 536–541. DOI: 10.18653/v1/2023.wassa-1.47.

[53] A. Numanoğlu, S. Ateş, N. Cicekli, and D. Küçük, "Empathify at WASSA 2024 empathy and personality shared task: Contextualizing empathy with a BERT-based context-aware approach for empathy detection," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 393–398.

[54] J. Mundra, R. Gupta, and S. Mukherjee, "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*, Online: Association for Computational Linguistics, Apr. 2021, pp. 112–116.

[55] T.-M. Lin, J.-Y. Chang, and L.-H. Lee, "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*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 548–552.

[56] T. Chavan, K. Deshpande, and S. Sonawane, "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*, Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 564–568.

[57] S. Buechel, A. Buffone, B. Slaff, L. Ungar, and J. Sedoc, "Modeling empathy and distress in reaction to news stories," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 4758–4765. DOI: 10.18653/v1/D18-1507.

[58] D. Omitaomu, S. Tafreshi, T. Liu, *et al.*, "Empathic conversations: A multi-level dataset of contextualized conversations," 2022.

[59] V. Barriere, S. Tafreshi, J. Sedoc, and S. Alqahtani, "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*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 214–227. DOI: 10.18653/v1/2022.wassa-1.20.

[60] P. Barros, N. Churamani, A. Lim, and S. Wermter, "The omg-empathy dataset: Evaluating the impact of affective behavior in storytelling," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2019, pp. 1–7. DOI: 10.1109/ACII.2019.8925530.

[61] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing," in *The Eleventh International Conference on Learning Representations, (ICLR)*, 2023.

[62] T. Wolf, L. Debut, V. Sanh, *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," 2020.

[63] S. Khorram, M. G. McInnis, and E. M. Provost, "Jointly aligning and predicting continuous emotion annotations," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1069–1083, 2019. DOI: 10.1109/TAFFC.2019.2917047.

[64] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 890–897. DOI: 10.1145/3123266.312338.

[65] B. T. Atmaja and M. Akagi, "Evaluation of error-and correlation-based loss functions for multitask learning dimensional speech emotion recognition," in *Journal of Physics: Conference Series*, IOP Publishing, vol. 1896, 2021, p. 012 004. DOI: 10.1088/1742-6596/1896/1/012004.

[66] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. DOI: 10.1016/0377-0427(87)90125-7.

[67] R. Chevi and A. Aji, "Daisy at WASSA 2024 empathy and personality shared task: A quick exploration on emotional pattern of empathy and distress," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 420–424.

[68] P. Pereira, H. Moniz, and J. P. Carvalho, "ConText at WASSA 2024 empathy and personality shared task: History-dependent embedding utterance representations for empathy and emotion prediction in conversations," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 448–453.

[69] S. Ouyang, J. M. Zhang, M. Harman, and M. Wang, "An empirical study of the non-determinism of chatgpt in code generation," *ACM Transactions on Software Engineering and Methodology*, Sep. 2024. DOI: 10.1145/3697010.

[70] K. Krippendorff, "Reliability," in *Content analysis: An introduction to its methodology*. Sage Publications, 2019. DOI: 10.4135/9781071878781.

[71] J. Wei, X. Wang, D. Schuurmans, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, Curran Associates, Inc., 2022, pp. 24 824–24 837.