The Practical Impacts of Theoretical Constructs on Empathy Modeling

Allison Lahnala[1], **Charles Welch**[1], **David Jurgens**[2], **Lucie Flek**[3,4]
[1] McMaster University, [2]University of Michigan
[3]Lamarr Institute for Machine Learning and Artificial Intelligence
[4]Bonn-Aachen International Center for IT, University of Bonn
{lahnalaa,cwelch}@mcmaster.ca, jurgens@umich.edu, flek@bit.uni-bonn.de

Abstract

Conceptual operationalizations of empathy in NLP are varied, with some having specific behaviors and properties, while others are more abstract. How these variations relate to one another and capture properties of empathy observable in text remains unclear. To provide insight into this question, we analyze the transfer performance of empathy models adapted to empathy tasks with different theoretical groundings. We study (1) the dimensionality of empathy definitions, (2) the correspondence between the defined dimensions and measured/observed properties, and (3) the conduciveness of the data to represent them, finding they have a significant impact to performance compared to other transfer setting features. Characterizing the theoretical grounding of empathy tasks as *direct*, *abstract*, or *adjacent* further indicates that tasks that directly predict specified empathy components have higher transferability. Our work provides empirical evidence for the need for precise and multidimensional empathy operationalizations.

1 Introduction

Empathy is considered a desirable aspect of interactions, but the difficulty of defining and operationalizing it inhibits scientific inquiry into the construct (Hall and Schwartz, 2019) and its role (e.g., in clinical encounters (Neumann et al., 2009)). Nevertheless, empathy is increasingly desired for conversational agents (Chen et al., 2023), leading many NLP researchers to investigate generative models of empathy, while appropriate evaluation frameworks are still limited (Lee et al., 2024). It is therefore a critical time to empirically investigate whether existing NLP operationalizations can provide a foundation for robust evaluation frameworks of empathic language, whether for examining empathy in generated output or human conversations.

NLP research on empathy detection and generation generally relies on the notion that facets of empathy may be measured or observed through language, whether they are considered a trait (Litvak et al., 2016; Yaden et al., 2024), a state (Buechel et al., 2018), or a social function in conversations (Omitaomu et al., 2022; Zhou and Jurgens, 2020). Empathy operationalizations in NLP range from singular ratings of empathy or emotion mirroring to fine-grained indicators and behaviors. Yet the underlying theoretical constructs and their mappings into language-based measurement tend to be underspecified, making it unclear whether our models effectively represent empathy and its components (Shetty et al., 2023). Given the often vague, varied use of "empathy" in NLP literature, it is not well understood how different operationalizations relate or what aspects of empathy they model (Lahnala et al., 2022).

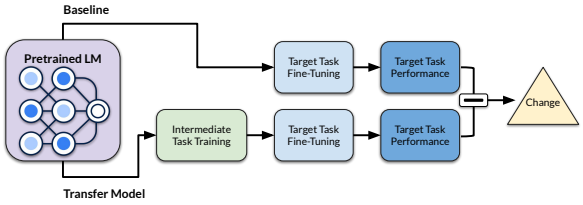

Figure 1: Experimental setup: How does intermediate task training on one empathy task impact performance on another target empathy task?

We aim to empirically quantify the practical effects of conceptual operationalizations to provide insight for developing future language-based empathy measurements and resources. We approach this with *transfer learning* experiments between 18 empathy tasks, hypothesizing that models trained on corpora with more similar constructs should provide more benefit to the target tasks (Alyafeai et al., 2020; Luo et al., 2022). Thus, transfer performance provides empirical indicators for answering our primary research question:

How does the theoretical grounding of empathy tasks impact transfer performance?

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 15571–15598
November 4-9, 2025

First, we assess the theoretical grounding of empathy operationalizations according to their definition *granularity*, *correspondence* between the operationalized measurements or observations to the definition, and *conduciveness* of the language situation to represent empathy in the data, along 5-point Likert scales. We identify three themes of the selected tasks as predicting aspects of the conceptualization *directly*, an *abstract* representation of empathy, or a construct *adjacent* to empathy. Then, each task is tested as an *intermediate* task for each other task as a *target* task in the transfer model setup shown in Figure 1, compared to a model without intermediate training.

We examine impacts to target task performance by considering the transfer settings between all task pairs, and which aspects of the settings are most predictive of performance change. Comparing the theoretical grounding assessments to predictive heuristics of transfer performance from prior work (e.g., dataset and task embedding similarity) (Poth et al., 2021) and other practical aspects of performance settings (e.g., sample and vocabulary sizes), we find that definition granularity and operationalization correspondence are significantly more predictive of transfer performance than these other features. Further, we analyze transfer performance according to the qualitative themes, finding that *direct* empathy tasks more often improved and less often harmed target tasks.

Our findings empirically confirm the importance of multidimensional operationalizations of empathy in NLP research (Lahnala et al., 2022). The lack of transferability, and moreover, the strength of the conceptual and measured/observed dimensionality in empathy tasks in predicting transfer performance highlight an important issue: abstract notions of empathy are insufficient for the development of NLP models. The lack of evidence that existing empathy prediction models are effective for downstream domains and language settings has important implications for developing robust evaluation frameworks for generative empathy models (Lee et al., 2024) and emotional intelligence in LLMs (Huang et al., 2023; Wang et al., 2023b; Sabour et al., 2024).

This paper contributes the first empirical analysis of the effects of construct operationalization on empathy prediction. Our methodology addresses the difficulty of disentangling aspects of construct operationalization from other transfer aspects, and therefore could be applied to investigate operationalizations of other complex constructs studied in NLP. We will publicly release the models developed for this work upon publication, offering additional tools for the analysis and evaluation of empathy in language. By integrating these models into a singular pipeline, we facilitate further investigation into the unique properties they capture, thereby advancing the development of a robust evaluation framework for empathy in language.

2 Background and Related Work

The motivations for NLP models of empathy are broad, ranging from developing empathic conversational agents (Morris et al., 2018; Rashkin et al., 2019; Naous et al., 2021), integration in counselor training tools (Tanana et al., 2016), analysis of interactions with support seekers (Xiao et al., 2015), and investigations of social interactions in online forums (Sharma et al., 2020; Zhou and Jurgens, 2020; Lahnala et al., 2021) which could support ongoing efforts investigating online support platforms, or "emotional first aid" (Weisberg et al., 2023). Moreover, empathic and emotional support capabilities are increasingly assessed in evaluation frameworks for LLMs (Huang et al., 2023; Chang et al., 2024; Manzoor et al., 2024). Several datasets including Rashkin et al.'s (2019) EmpatheticDialogs for generation, Sharma et al.'s (2020) Epitome, Buechel et al.'s (2018) NewsReactions, among others, have been introduced for developing models for empathy detection and generation, with much interest focused on the latter. These systems often incorporate a module for sentiment or emotion (e.g., Majumder et al.; Rashkin et al.; Lin et al.), and some aim to incorporate common sense reasoning or other grounding knowledge (e.g., Sabour et al.; Li et al.; Lee et al.).

Despite broad motivations for NLP models of empathy, numerous works have pointed out the limitations due to limited and/or unclear conceptual operationalizations of empathy. Research on detecting or generating "empathic language," often does not specify empathic processes, behaviors, or social expectations of empathic expression (Lahnala et al., 2022). Research tends to center on emotional empathy, leaving other facets underexplored, such as how cognitive empathic processes may be observed in interactions (Sharma et al., 2020). Precise definitions and measurement approaches of empathy in language remain open challenges (Shetty et al., 2023); measurements tend to abstract the

Dataset	Samples	$\|V\|$	Total Tokens
Condolence	1,004	15,139	353,951
Conv	11,176	11,040	3,755,038
EmpDial	76,494	26,054	2,642,492
EmpDial EI	53,414	26,447	5,043,441
EmpQT	20,201	16,703	1,602,514
Empathy & Hope	1,282	6,461	31,415
Epitome	3,023	11,882	324,084
MI	15,645	14,876	3,344,011
News Stories	2,700	6,324	228,764

Table 1: Number of samples, size of the vocabulary ($|V|$), and total number of tokens in each dataset.

construct into single ratings, which fails to evaluate important facets of empathy (Lee et al., 2024).

To investigate the impacts of these issues empirically, our study explores the transferability between empathy tasks. Our study draws on Poth et al. (2021), which considered heuristics to estimate task transferability between a variety of intermediate and target NLP tasks. They found that the dataset size and task similarity are often reasonable predictors of transfer performance. They average sentence-BERT embeddings over a corpus and measure similarity to predict transfer performance (Reimers and Gurevych, 2019). Other works, such as Luo et al. (2022) examine both neural network (computer) models and fMRI data to derive task representations. Vu et al. (2020) explore using task embeddings computed from the Fisher information matrix to predict transfer performance, as the information about useful model parameters serves as a "rich source of knowledge about the task itself." These works have demonstrated that the similarity of these embeddings between tasks is an effective heuristic for estimating transferability. We use these heuristics in addition to other practical factors that can affect transfer performance (e.g., the number of samples, tokens, and data sources) to assess how relevant the theoretical grounding is to transfer performance by comparison.

3 Dataset and Task Descriptions

We study 18 regression and classification tasks across nine English-language datasets (Table 1), with various empathy constructs underlying the dataset collection and task definitions. We outline them here, and provide details in Appendix A.

Social Media Datasets. The Motivational Interviewing (MI) (Welivita and Pu, 2022), Condolence (Zhou and Jurgens, 2020), and Epitome (EPIT) (Sharma et al., 2020) datasets are sourced from Reddit. MI is annotated with MI counselor behaviors at the sentence level using the aforementioned Motivational Interviewing Treatment Integrity (MITI) scheme (Moyers et al., 2003) in interactions in mental health forums. Epitome also contains interactions in mental health forums, labeled based on the degree to which certain empathy communication mechanisms are exhibited in peer-supporter responses. The Condolence dataset (Zhou and Jurgens, 2020) contains exchanges from Reddit forums with empathy ratings based on an appraisal theory of empathy with six dimensions (Lamm et al., 2007; Wondra and Ellsworth, 2015). The Empathy Hope dataset (Yoo et al., 2021) contains geopolitical tweets from India and Pakistan to classify supportive content that expresses empathy, distress, or solidarity.

Crowd-sourced Datasets. The News dataset (Buechel et al., 2018) contains empathetic essay reactions to news articles with first-person empathy measurements, and continuous empathy and distress scores are derived from Batson's Empathic Concern – Personal Distress Scale (Batson et al., 1987). The Conv dataset (Omitaomu et al., 2022) contains crowd-sourced conversations between two participants about articles in the News dataset, with turn-level annotations of emotional polarity, emotional intensity, and empathy. The EmpDial dataset (Rashkin et al., 2019) contains dialogues grounded on specific emotions, drawing on emotion theories such as Ekman's (Ekman, 1971, 1992) and Plutchik's (Plutchik, 1984). EmpDial EI (Welivita and Pu, 2020) and EmpDial QI (Svikhnushina et al., 2022) are subsets of the EmpDial dataset annotated according to taxonomies of empathic intents in responses and questions.

4 Empathy Operationalizations

To study how conceptual operationalizations of empathy tasks impact transfer performance, we assess empathy tasks within NLP based on core premises underlying their scientific methods, namely, 1) the construct *Definition*, 2) the *correspondence (Link)* of the measurements and observations used to operationalize the construct to the construct itself, and 3) the *Conduciveness* of the data to represent the construct. We developed an annotation task, defining criteria for these aspects along a 5-point Likert scale and rate each task (discussed in §4.1; full instructions and scoring criteria are in Figures 9 and 10 in the Appendix).

	Def.	Link	Cond.	Mean
Direct				
News Empathy	5.0	4.5	4.0	4.50
Conv Empathy	5.0	3.5	5.0	4.50
EmpDial EI.	4.5	5.0	3.0	4.17
News Distress	4.5	4.0	4.0	4.17
EmpDial QInt.	4.0	5.0	3.0	4.00
MI Behavior	3.5	4.0	3.5	3.67
Abstract				
Condolence	5.0	2.0	3.5	3.50
Epitome EX	3.0	3.0	3.5	3.17
Epitome ER	3.0	3.0	3.5	3.17
Epitome IP	3.0	3.0	3.5	3.17
Empathy Hope*	2.5	2.5	1.0	2.00
MI Adherent*	2.5	2.5	3.0	2.67
Adjacent				
EmpDial QAct.*	3.0	4.0	3.0	3.33
Conv EmoPol*	3.0	2.0	4.5	3.17
News Emotion	3.0	2.5	3.5	3.00
Conv EmoInt	3.0	1.5	3.5	2.67
EmpDial Emo	2.0	1.0	2.0	1.67
EmpDial Role	1.5	1.0	1.5	1.33

Table 2: **Themes of Theoretical Grounding.** The means of the annotators' ratings per task are shown on the right. We see that the mean ratings almost separate the datasets into the three distinct groups, which map to the authors deliberative categorization. Deviations are marked with *.

In addition, we carefully considered and discussed the empathy tasks and identified three broad themes (discussed in §4.2) regarding the relationship between the empathy tasks and their theoretical grounding: 1) *direct* empathy, 2) empathy *abstraction*, and 3) empathy *adjacent* tasks. Both the annotation of operationalizations and qualitative categorization into themes enable a qualitative analysis of the impact of each on transfer performance.

4.1 Rating Conceptual Operationalizations

The rating criteria for the definition and correspondence aspects are informed by the empathy definition and evaluation themes identified by Lahnala et al. (2022), and prior work in social psychology (Hall and Schwartz, 2019), which similarly surveyed the diversity of conceptual and operational definitions of empathy and the correspondence between the measurements and construct. The third component relates to contextualizing language behaviors within specific situations or contexts (Boyd and Markowitz, 2024), and considering expectations around empathy that may influence the data.

Construct Definition. We assess the granularity of each task's conceptual definition of empathy, considering the number of dimensions, behaviors, and other details about the construct provided by the works. They range from providing *no* conceptual definition to embracing empathy's multidimensionality with descriptions of many dimensions, factors, and characteristics of empathic experiences, interactions, language, and conversational behaviors.

Operational Correspondence (Link). We assess the correspondence of the operationalization of empathy to the defined construct (referred to as *Link* for short) by considering their measurements and observations. For the NLP context, these are typically reflected by the task labels; how directly are the task labels connected to the defined aspects of the underlying construct? Higher ratings reflect tasks that cover several possible empathic behaviors or indicators of the empathy construct. In contrast, lower ratings reflect tasks with labels that are an abstraction of the empathy construct or no labels relating to the construct are involved. As an example, the development of the *Condolence* dataset is based on a fine-grained construct, an *appraisal theory* of empathy (Lamm et al., 2007; Wondra and Ellsworth, 2015) which specifies six perspective dimensions along which an observer can appraise a target's situation. However, the task labels are singular ratings of empathy that abstract the underlying construct based on trained annotators' assessment of how the empathic observer's appraisals align with the support-seeking target by considering the six dimensions at once. Thus, *Condolence* is rated highly by definition but low by link.

Conduciveness. We assess the Conduciveness of the *language scenario* to capture empathic processes in the interactions; i.e., how much could one expect *empathy* to be observable in the language? However, any natural interaction between humans could represent empathic processes (Debnath and Conlan, 2023). Thus, we base the ratings on plausible expectations or assumptions that can be made about the scenario, considering aspects like the social norms of the scenario and the data collection methods. For example, we may consider whether a certain level of common ground between participants can be assumed based on experimental controls. Social norm considerations deal with the nature of the language or conversation scenario; for instance, empathy is a social norm for mental health support and therapy conversations, but it is not an expectation of scientific articles. For ex-

ample, both annotators rate *Conv Empathy* highly based on aspects of the experimental design (see Table 3). Before having a conversation with each other, two participants independently performed the same tasks intended to induce empathic processes/experiences. First, they read the same inherently emotional news article, which grounds their knowledge of the conversation topic; then, they are asked to describe their empathic experience in a short essay. These facets of the pre-conversation task imply the participants begin the conversation with common ground and that they have performed more deliberative empathic processes beyond initial emotional reactions, which we consider conducive to cognitive empathy.

Annotation Task. Two authors of this study completed the task and are knowledgeable about empathy constructs within and outside the NLP field. All ratings collected from the annotation task are reported in Table 3 in the Appendix. Table 4 shows the inter-rater agreement measured by Krippendorff's α and the Spearman rank correlation coefficients r and p-values between the annotators on each aspect. The agreement and correlation measures are highest on the Definition and Link. The higher agreements on Definition and Link versus the lower agreement on Conduciveness are likely affected by relatively more objectivity in the case of Definition and Link. In contrast, Conduciveness is less defined and more subjective to interpretation regarding several aspects that could describe a "language scenario." Figure 4 shows the degree of difference between annotator ratings on each aspect. We observe that the Conduciveness ratings more often differ by one point; in four cases, they differ by two points. Specific disagreements can be observed in Table 3.

4.2 Themes of Theoretical Grounding

Table 2 shows the characterization of each task according to the themes described in this section.

Direct empathy tasks generally have fine-grained construct definitions that link directly to the task labels. This group includes, for instance, MI Behavior, which directly labels counseling behaviors defined by the MITI construct, and EmpDial QInt and EI, which involve the labels of the empathic intents that *comprise* their construct, directly at the dialogue act level. This group also includes the News Empathy and Distress (Buechel et al., 2018), though News Distress borders on belonging to the next group, *empathy abstraction*. The labels are ratings that capture two facets of an empathy construct, *distress* and *empathy* (empathic concern) are the primary facets they target, so we include them here. Furthermore, the ratings are empirically derived by a *multi-item* scale that measures the empathic observers' internal empathic experience upon reading empathy-inducing articles; thus, we consider them more directly connected to the construct than a third person's abstractive assessment. Similarly, Conv Empathy measures the empathy construct with the same data but through the assessment of a conversation, rather than essay response.

Empathy abstraction tasks range in granularity of the construct definitions but involve abstraction of the underlying construct, for instance, by binarizing or rating the construct more broadly. This includes the *Condolence* task, which abstracts a fine-grained construct as discussed in Section 4.1, and the *MI Adherent* task, which is a binary abstraction of a set of counseling behaviors related but not explicitly connected to a set of empathic process components. Empathy Hope is included here because it contains labels of supportive content, which is broad and may contain empathy. The three Epitome tasks do separate cognitive and emotional aspects of empathy, but rely on third-person annotations of the text that use strong/weak/none, so we consider it as an abstraction of the expression of each aspect.

Empathy-adjacent tasks as reflected by the name are those that involve predicting labels that reflect concepts that *relate* to empathy, rather than labels connected to the empathy construct. Naturally, these have a lower level of empathy construct granularity and Link to empathy construct, but are useful aspects to analyze in empathic interactions, such as emotions and emotion intensity reflected in EmpDial Emo, Conv EmoPol, Conv EmoInt and News Emotion, and dialogue roles and acts reflected in EmpDial Role and EmpDial QAct.

5 Experiments: Empathy Task Transferal

Our study aims to identify the ways that conceptual operationalizations of empathy impact transferability between models. Our experimental paradigm shown in Figure 1 is to compare models baseline *target task* models to models tuned first on an *intermediate task* and adapted for the *target* task. Testing each of the 18 empathy tasks as an intermediate for each of the other 17 tasks as targets provides a large space for examining how factors of the trans-

fer settings relate to performance differences from the baselines.

Full fine-tuning for this number of experiments would require substantial computational resources and energy (Wang et al., 2023a). Therefore, we use the parameter-efficient adapter-tuning strategy. Adapters are a lightweight tuning strategy in which the pre-trained model parameters are frozen, and only the weights of new layers injected into the model are updated when training for a downstream task; this strategy has often matched the performance of full fine-tuning (Houlsby et al., 2019). Furthermore, this is the same approach used by Poth et al. (2021) for exploring heuristics for estimating task transferability between numerous NLP tasks, which informs our work.

All models have a Transformer-based architecture (Vaswani et al., 2017) with adapter modules tuned for the empathy tasks. We use RoBERTa-base (Liu et al., 2019) as the pre-trained Transformer. We tested several types of adapters during a hyperparameter search, finding best outcomes from single bottleneck adapters (Pfeiffer et al., 2020) followed by LoRA (Hu et al., 2022), and therefore use the bottleneck adapters across all experiments. Full training and hyperparameter search details are provided in Appendix C. We encode the prior post or turn and the target utterance for each dataset that includes context (all but News and Empathy Hope).

Baselines. We obtain baseline predictions for each target task by training a task adapter with a prediction head and obtaining predictions from this model on the test split. We performed a hyperparameter search and used the best configuration for each task, shown in Table 5.

Empathy-to-Empathy task transfer. The 18 task adapters trained for the target task baselines are subsequently used in the transfer experiments as *intermediate task* adapters. To adapt them for the remaining 17 target tasks, we remove the prediction head from the intermediate adapter, and add a new task adapter and prediction head for the target task, using the stacked composition setup from Pfeiffer et al. (2020). We train this composition for the target task and use the resulting model for inference on the target test dataset.

Results Overview. The F1 scores for the classification tasks are provided in Table 7 and the Pearson r scores for regression tasks are in Table 8. For each intermediate-to-target pair, we compute the percent change in performance compared to the baselines. The values for the classification target tasks are shown in a heatmap in Figure 5 and for the regression target tasks in Figure 6. Statistically significant performance gains across tasks are relatively rare in our experimental results, indicating little is gained from intermediate empathy task tuning. The following sections are in-depth examinations of the results with respect to our RQ.

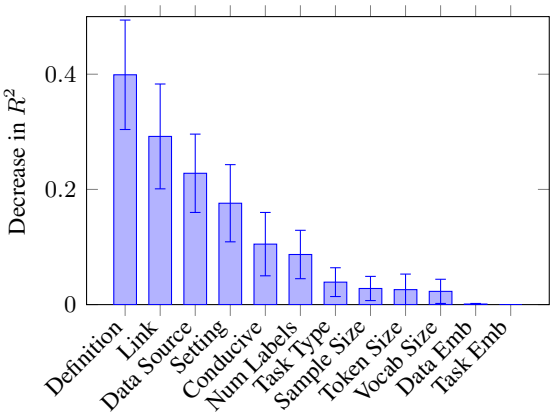

Figure 2: Feature importances in regression fit to improvement over baseline calculated as the difference in R^2 when permuting the feature.

5.1 Features Impacting Transferability

To understand *how* conceptual operationalizations of empathy impact transfer performance, we consider the importance of the definition granularity, operational correspondence, and data conduciveness (§4.1) in predicting the performance change between the baseline and transfer models, in comparison to other features of the transfer settings.

Basic Features: For each intermediate-target task pair, we consider whether the following increases, decreases, or remains the same: the i) number of data samples, ii) number of tokens, iii) number of unique tokens (vocabulary), and (iv) number of labels. Then, we consider (v) the task types–whether the transfer goes from classification to classification or regression, and vice versa; (vi) the source and target tasks' data source (e.g., Reddit-to-Twitter); and (vii) language setting (e.g., conversation, social media, or essay).

Transferability Heuristics: We also consider the similarities between the pairs' (viii) dataset embeddings and (ix) task embeddings, which were found be effective at estimating the transferability between NLP tasks (Vu et al., 2020; Poth et al., 2021; Sileo and Moens, 2022), with higher similarities being associated with better transfer performance.

Following Poth et al. (2021)'s approach, we use the `all-mpnet-base-v2` model to compute Sentence BERT (SBERT) embeddings over each dataset and subsequently the cosine similarity between each dataset's embedding. We follow the task embedding similarity aggregation from previous work (Vu et al., 2020), which combines BERT components using reciprocal rank fusion (Cormack et al., 2009). Heatmaps of the similarities are shown in Figures 7 and 8.

Construct Features: These features are the differences between the intermediate and target task ratings on the annotated aspects.

Quantifying Feature Importance: We fit a support vector regression model (Awad et al., 2015) on these features to predict the percentage improvement by the transfer models over the baselines. We measure feature importance using the permutation approach, whereby each feature is separately permuted across instances, and the impact on R^2 is measured (Breiman, 2001). We permute 50 times for each of the ten models trained on random train/test splits of the 306 trials.

Figure 2 shows the feature importances averaged across permutations.[1] Definition granularity and operational correspondence (link) are the most predictive of transfer performance, followed by the data source and language setting. Conduciveness and the remaining basic features are relatively insignificant compared to these features. The heuristic features (data and task emb) were *not* predictive.

5.2 Qualitative Themes and Performance

The themes outlined in §4.2 provide researchers a quick and intuitive way to characterize the theoretical grounding of their empathy tasks: does the task *directly* predict certain components of the empathy construct, predict an *abstract* representation of the construct, or some other construct related or *adjacent* to empathy? Grouping tasks by these themes, we perform a qualitative analysis of how the theoretical groundings transfer to other ones.

Figure 3 (top) shows the count of adapters that significantly[2] outperformed and underperformed the task baselines aggregated by these themes. Direct empathy tasks as intermediates are the most frequent to outperform the baseline, but only for direct empathy and empathy adjacent to target tasks.

[1] Values are provided in Table 6.

[2] Significance indicates p-value < 0.05 provided by a non-parametric bootstrap test with resampling.

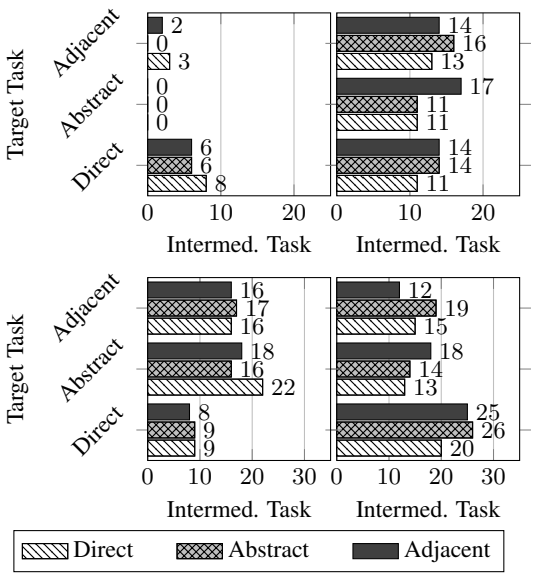

Figure 3: **Transfer performance by theme.** Top: Significant improvement (left) and significant harm (right) counts for full data. Bottom: Insignificant difference or significant improvement (left) and significant harm (right) counts for limited data.

Intermediate tuning does not benefit any abstract tasks. The transfer learning models often significantly *harm* the performance, which occurs least frequently with direct intermediate tasks. The direct empathy tasks are also the most frequent targets to *gain* from intermediate tuning.

5.3 Case Study: Limited Data

A motivation for studying empathy task transferability is the potential to utilize empathy models trained on existing resources to increase their scale when training data is limited, and support the construction of new resources. For example, transferability has critical value for supporting technologies for health care like clinician communication training, where annotations are costly and data is sensitive (Tanana et al., 2019; Imel et al., 2017). While the previous section showed little value for *improving* target task performance, could intermediate tuning with *limited training data* at least achieve *similar* results as using no intermediate tuning but the *full training data*?

To explore this, we tested intermediate tasks in models that use *half* the amount of target task data for target task tuning and compared the performance differences to the target task-only baseline developed on the complete training set. For these results shown in Figure 3 (bottom), we are interested in the rate of significant losses compared to

the rate of insignificant differences (or significant improvements), as this indicates whether the transfer effectively enables similar performance to a scenario in which we had twice the training data. As intermediate tasks, the task themes have similar rates of insignificant differences or improvements, only direct empathy tasks showing more benefits to abstract target tasks. However, the direct tasks as the target task are much less likely to benefit, shown also by significant harms.

6 Discussion

Conceptual operationalizations have practical impacts. We examined the theoretical grounding of empathy tasks according to the definition granularity, the correspondence between operationalization and defined components, and the conduciveness of the data to reflect empathy. Analyzing these features in a model fit to the change in performance from baseline to transfer model revealed the significant influence conceptual operationalizations have over transfer performance. This underscores the importance of well-defined, multidimensional constructs and measurements/observed that directly correspond to them (§5.1).

Constructs are not equally helpful or harmful. We explored how tasks function as intermediates and targets based on the themes defined in §5.2, and whether using the transfer models could support limited data settings (§5.3). While few transfer models overall significantly improve performance, abstraction target tasks are never improved, indicating that different theoretical empathy constructs are unable to map meaningful properties to abstract ones. Conversely, abstract and adjacent tasks may map to specific components of direct tasks. However, these observations are overshadowed by the much higher rates of performance harms observed for all themes as intermediates and targets, occurring slightly less often from direct intermediates. The direct tasks had the most frequent significant improvement and least performance losses compared to baselines fine-tuned only on the target task. The limited data setting demonstrated direct empathy had less to gain and more to lose from intermediate transfer. Small differences between the groups may suggest that direct tasks are more likely to effectively support resource expansion and construction, but there is not substantial evidence.

NLP needs precise, multidimensional empathy constructs. The main takeaway of our findings is that fine-grained definitions operationalized via measurable or observable characteristics in language for a construct such as empathy, which is not directly measurable, are necessary for more reliable representations. The tasks with better transferability may be more beneficial, for instance, when deciding which training data to supply to the training of an empathetic model, but compelling evidence of the benefits is yet to be shown. Further researchers of NLP empathy modeling should aim for well-defined, multidimensional constructs, and use measurements/observations that directly correspond to the specific components of the construct. Future work could draw inspiration from approaches in psychometrics that utilize multi-item instruments to measure latent constructs (El-Den et al., 2020), as was done in developing the *News* and *Conv* tasks (Buechel et al., 2018; Omitaomu et al., 2022).

Empathy constructs studied in the field of psychology commonly involve broad components such as cognitive and emotional empathy, yielding a variety of measurement instruments depending on what facet of empathy is the target of study (Cuff et al., 2016). Some measures may estimate internal empathic experiences of an empathic observer (e.g., Batson's Empathic Concern – Personal Distress Scale (Batson et al., 1987) used by Buechel et al., 2018), while others may characterize empathy in dialog with discourse frameworks (e.g., Pounds, 2011; Rey Velasco et al., 2022). While research on empathy in psychology and other fields could provide a foundation for NLP research, these fields face many of the same operationalization challenges (Coll et al., 2017; Hall and Schwartz, 2019). Hall and Schwartz's (2019) survey of psychology literature highlights a prevalent lack of consistency in definitions and measurements, validations of operationalizations, and frequently weak correspondence between construct definition and measurement. While they also observed the rates of stating conceptual definitions and referring to empathy's multidimensionality increasing over time, they argue that the application of the term "empathy" could be refined to lower-level constructs and in many cases the term is inappropriate in a scientific sense for their research objective.

Ultimately, as empathy is not directly measurable, operationalizing requires defining multiple indicators and characteristics that can be measured, observed, and tested. As Hall and Schwartz put it, "The continued vague use of the term empathy to characterize a wide range of different methods and

definitions can only dilute the value of scholarship." We therefore invite further discussions regarding the question of whether the construct we intend to study is best described as "empathy" as opposed to a lower-level construct: What is the specific use for models that can detect or generate empathy? If we scrutinize the empathy construct against the purposes and goals of specific systems, are there more specific and operationalizable constructs that are more appropriate?

7 Conclusion

As interest in systems that model empathy grows, it is crucial to address the limitations of NLP research on empathy of whether the conceptual operationalizations lead to reliable models of empathy in language. This work leveraged an intermediate-task transfer experiment paradigm to investigate transferability of 18 empathy-related tasks in NLP research. We analyzed the performance on the basis of each task's conceptual operationalization of empathy to answer the research question of how the theoretical grounding of empathy tasks impacts empathy modeling. The critical finding is that the *granularity* of empathy definitions *and* how directly the measurements and observations in the data *correspond* to the defined components are significant factors of predicting transfer performance. Future researchers should carefully construct empathy modeling tasks according to well-defined constructs that are measurable or observable in the language data. The code for our experiments is available at `https://github.com/alahnala/empathy-transferability`.

Limitations

For this study, we chose widely used datasets to provide insight into the properties of resources that researchers are currently exploring, as well as a recently introduced large-scale dataset that differs more in terms of the type of tasks. Further work may explore transferable properties from other empathy and empathy-related datasets and task types, such as generation.

We note that we introduced three construct groups through deliberation with the authors of this paper. This process involved a few hours of discussion of each datasets qualities and measurements. This process is a qualitative research approach that not easily reproducible, but nonetheless provides a framework for thinking about empathy datasets in future work. Quantitatively, we showed that our annotation process can be used to arrive at a similar separation of groups using the average of annotated aspects (see Table 2, however these three numbers by themselves may not capture the full picture. This is a starting point in a larger discussion about how to categorize the constructs of empathy and better understand their impact on downstream task performance. Future work should seek more concrete validation of the construct groups through refining the annotation process.

Our study only contains English data and, thus, only captures properties of how empathy is expressed in English. We expect that there are socio-variations in how empathy is expressed and perceived, not only across languages but also across different English speakers, contexts, and relationships between speakers. Future work could explore transfer learning to support empathy modeling in multi-lingual spaces. There are a variety of empathy resources for other languages, such as Arabic (Naous et al., 2021), Italian (Alam et al., 2018; Sanguinetti et al., 2020), Japanese (Ito et al., 2020), German (Wambsganss et al., 2021), and Chinese (Sun et al., 2021). This work could explore zero-shot or few-shot cross-lingual transfer to languages without annotated empathy resources, as was done by Pfeiffer et al. (2020) for other NLP tasks. In our experiments, we test transferring from one task to another in a stacked composition adapter setup. Further research may consider combinations of multiple-task adapters and multi-task learning.

Acknowledgments

This work was partially supported by the BMFTR, the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence and the National Science Foundation through Grant No. IIS-2143529. We also thank the reviewers for their invaluable comments, which helped strengthen the quality of this work.

References

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

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. *ArXiv preprint*, abs/2007.04239.

Mariette Awad, Rahul Khanna, Mariette Awad, and Rahul Khanna. 2015. Support vector regression. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, pages 67–80.

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.

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).

Ryan L. Boyd and David M. Markowitz. 2024. Verbal behavior and the future of social science. *American Psychologist*. Place: US Publisher: American Psychological Association.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

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.

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

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

Michel-Pierre Coll, Essi Viding, Markus Rütgen, Giorgia Silani, Claus Lamm, Caroline Catmur, and Geoffrey Bird. 2017. Are we really measuring empathy? proposal for a new measurement framework. *Neuroscience & Biobehavioral Reviews*, 83:132–139.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759.

Benjamin M.P. Cuff, Sarah J. Brown, Laura Taylor, and Douglas J. Howat. 2016. Empathy: A review of the concept. *Emotion Review*, 8(2):144–153.

Alok Debnath and Owen Conlan. 2023. A critical analysis of empatheticdialogues as a corpus for empathetic engagement. In *Proceedings of the 2nd Empathy-Centric Design Workshop*, EMPATHICH '23, New York, NY, USA. Association for Computing Machinery.

Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Sarira El-Den, Carl Schneider, Ardalan Mirzaei, and Stephen Carter. 2020. How to measure a latent construct: Psychometric principles for the development and validation of measurement instruments. *International Journal of Pharmacy Practice*, 28(4):326–336.

N. J. Enfield, Tanya Stivers, and Stephen C. Levinson. 2010. Question–response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics*, 42(10):2615–2619.

Alice F Freed. 1994. The form and function of questions in informal dyadic conversation. *Journal of pragmatics*, 21(6):621–644.

Alison Gopnik and Henry M Wellman. 2012. Reconstructing constructivism: Causal models, bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6):1085.

Robert M. Gordon. 1992. The simulation theory: Objections and misconceptions.

James J Gross. 2013. *Handbook of emotion regulation*. Guilford publications.

Judith A Hall and Rachel Schwartz. 2019. Empathy present and future. *The Journal of social psychology*, 159(3):225–243.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR, 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*.

Karen Huang, Michael Yeomans, Alison Wood Brooks, Julia Minson, and Francesca Gino. 2017. It doesn't hurt to ask: Question-asking increases liking. *Journal of personality and social psychology*, 113(3):430.

Zac E Imel, Derek D Caperton, Michael Tanana, and David C Atkins. 2017. Technology-enhanced human interaction in psychotherapy. *Journal of counseling psychology*, 64(4):385.

Koichiro Ito, Masaki Murata, Tomohiro Ohno, and Shigeki Matsubara. 2020. Relation between degree of empathy for narrative speech and type of responsive utterance in attentive listening. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 696–701, Marseille, France. European Language Resources Association.

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

Allison Lahnala, Yuntian Zhao, Charles Welch, Jonathan K. Kummerfeld, Lawrence C An, Kenneth Resnicow, Rada Mihalcea, and Verónica Pérez-Rosas. 2021. Exploring self-identified counseling expertise in online support forums. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4467–4480, Online. Association for Computational Linguistics.

Claus Lamm, C. Daniel Batson, and Jean Decety. 2007. The neural substrate of human empathy: Effects of perspective-taking and cognitive appraisal. *Journal of Cognitive Neuroscience*, 19(1).

Andrew Lee, Jonathan Kummerfeld, Larry An, and Rada Mihalcea. 2023. Empathy identification systems are not accurately accounting for context. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1686–1695, Dubrovnik, Croatia. Association for Computational Linguistics.

Andrew Lee, Jonathan Kummerfeld, Larry Ann, and Rada Mihalcea. 2024. A comparative multidimensional analysis of empathetic systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–189, St. Julian's, Malta. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI, 2022, Thirty-Fourth Conference on Innovative Applications of Artificial, Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances, in Artificial Intelligence, EAAI 2022 Virtual Event, February 22, - March 1, 2022*, pages 10993–11001. AAAI Press.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Marina Litvak, Jahna Otterbacher, Chee Siang Ang, and David Atkins. 2016. Social and linguistic behavior and its correlation to trait empathy. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 128–137, Osaka, Japan. The COLING 2016 Organizing Committee.

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*, abs/1907.11692.

Sarah Peregrine Lord, Elisa Sheng, Zac E Imel, John Baer, and David C Atkins. 2015. More than reflections: Empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, 46(3):296–303.

Yifei Luo, Minghui Xu, and Deyi Xiong. 2022. CogTaskonomy: Cognitively inspired task taxonomy is beneficial to transfer learning in NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 904–920, Dublin, Ireland. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 1022–1035.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh,

Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Muhammad Arslan Manzoor, Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. Can machines resonate with humans? evaluating the emotional and empathic comprehension of lms. *arXiv preprint arXiv:2406.11250*.

Jim McCambridge, Maria Day, Bonnita A Thomas, and John Strang. 2011. Fidelity to motivational interviewing and subsequent cannabis cessation among adolescents. *Addictive behaviors*, 36(7):749–754.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *Journal of medical Internet research*, 20(6):e10148.

Theresa B Moyers, Tim Martin, Jennifer K Manuel, William R Miller, and D Ernst. 2003. The motivational interviewing treatment integrity (MITI) code: Version 2.0. *Unpublished manuscript. Albuquerque, NM: University of New Mexico, Center on Alcoholism, Substance Abuse and Addictions*.

Theresa B Moyers and William R Miller. 2013. Is low therapist empathy toxic? *Psychology of Addictive Behaviors*, 27(3):878.

Tarek Naous, Wissam Antoun, Reem Mahmoud, and Hazem Hajj. 2021. Empathetic BERT2BERT conversational model: Learning Arabic language generation with little data. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 164–172, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Melanie Neumann, Jozien Bensing, Stewart Mercer, Nicole Ernstmann, Oliver Ommen, and Holger Pfaff. 2009. Analyzing the "nature" and "specific effectiveness" of clinical empathy: A theoretical overview and contribution towards a theory-based research agenda. *Patient Education and Counseling*, 74(3):339–346. Theories in Health Communication Research.

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*, abs/2205.12698.

Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. 2020. Hope speech detection: A computational analysis of the voice of peace. In *ECAI 2020*, pages 1881–1889. IOS Press.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435, Vancouver, Canada. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4.

Kathryn I Pollak, Stewart C Alexander, James A Tulsky, Pauline Lyna, Cynthia J Coffman, Rowena J Dolor, Pål Gulbrandsen, and Truls Østbye. 2011. Physician empathy and listening: Associations with patient satisfaction and autonomy. *The Journal of the American Board of Family Medicine*, 24(6):665–672.

Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. What to pre-train on? Efficient intermediate task selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gabrina Pounds. 2011. Empathy as 'appraisal': A new language-based approach to the exploration of clinical empathy. *Journal of Applied Linguistics and Professional Practice*, 7(2):139–162.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

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

Elena Rey Velasco, Hanne Sæderup Pedersen, Timothy Skinner, et al. 2022. Analysis of patient cues in asynchronous health interactions: Pilot study combining empathy appraisal and systemic functional linguistics. *JMIR Formative Research*, 6(12):e40058.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Juanzi

Li, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *CoRR*, abs/2402.12071.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: Commonsense-Aware empathetic response generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI, 2022, Thirty-Fourth Conference on Innovative Applications of Artificial, Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances, in Artificial Intelligence, EAAI 2022 Virtual Event, February 22, - March 1, 2022*, pages 11229–11237. AAAI Press.

Manuela Sanguinetti, Alessandro Mazzei, Viviana Patti, Marco Scalerandi, Dario Mana, and Rossana Simeoni. 2020. Annotating errors and emotions in human-chatbot interactions in Italian. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 148–159, Barcelona, Spain. Association for Computational Linguistics.

Robert L Selman. 1980. *The growth of interpersonal understanding: Developmental and clinical analyses*. Academy Press.

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

Vishal Anand Shetty, Shauna Durbin, Meghan S Weyrich, Airín Denise Martínez, Jing Qian, and David L Chin. 2023. A scoping review of empathy recognition in text using natural language processing. *Journal of the American Medical Informatics Association*, page ocad229.

Damien Sileo and Marie-Francine Moens. 2022. Analysis and prediction of NLP models via task embeddings. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 633–647, Marseille, France. European Language Resources Association.

Tania Singer and Olga M Klimecki. 2014. Empathy and compassion. *Current biology*, 24(18):R875–R878.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. PsyQA: A Chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503, Online. Association for Computational Linguistics.

Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. 2022. A taxonomy of empathetic questions in social dialogs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2952–2973, Dublin, Ireland. 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.

Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment*, 65:43–50.

Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *Journal of medical Internet research*, 21(7):e12529.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2021. Supporting cognitive and emotional empathic writing of students. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4063–4077, Online. Association for Computational Linguistics.

Xiaorong Wang, Clara Na, Emma Strubell, Sorelle Friedler, and Sasha Luccioni. 2023a. Energy and carbon considerations of fine-tuning BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9058–9069, Singapore. Association for Computational Linguistics.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023b. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Ori Weisberg, Shiri Daniels, and Eran Bar-Kalifa. 2023. Emotional expression and empathy in an online peer support platform. *Journal of Counseling Psychology*, 70(6):671.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2022. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Joshua D Wondra and Phoebe C Ellsworth. 2015. An appraisal theory of empathy and other vicarious emotional experiences. *Psychological review*, 122(3):411.

Bo Xiao, Zac E Imel, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2015. "rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PloS one*, 10(12):e0143055.

David B Yaden, Salvatore Giorgi, Matthew Jordan, Anneke Buffone, Johannes C Eichstaedt, H Andrew Schwartz, Lyle Ungar, and Paul Bloom. 2024. Characterizing empathy and compassion using computational linguistic analysis. *Emotion*, 24(1):106.

Clay H. Yoo, Shriphani Palakodety, Rupak Sarkar, and Ashiqur KhudaBukhsh. 2021. Empathy and hope: Resource transfer to model inter-country social media dynamics. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 125–134, Online. Association for Computational Linguistics.

Aston Zhang, Yi Tay, Shuai Zhang, Alvin Chan, Anh Tuan Luu, Siu Cheung Hui, and Jie Fu. 2021. Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with 1/n parameters. In *9th International Conference on Learning Representations, ICLR 2021 Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Naitian Zhou and David Jurgens. 2020. Condolence and empathy in online communities. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 609–626, Online. Association for Computational Linguistics.

Dolf Zillmann. 2008. *Empathy Theory*. John Wiley & Sons, Ltd.

A Dataset and Task Descriptions

This appendix provides the details of each dataset and task. Tables 9 through 12 contain splits and label distributions.

A.1 Condolence

The Condolence Empathy dataset contains expressions of condolences exchanged on Reddit forums annotated with empathy ratings from 1.0 to 5.0 (Zhou and Jurgens, 2020). Two annotators were trained on an *appraisal theory* of empathy (Lamm et al., 2007; Wondra and Ellsworth, 2015), which specifies six perspective dimensions that an observer may appraise of the target: (1) pleasantness, (2) anticipated effort, (3) situational control, (4) who was responsible for the situation (self or other), (5) attentional activity, and (6) certainty about the situation or its aftermath. The annotators assessed the observers' condolences in response to targets' expressing their situation, basing their single rating on consideration of these dimensions. The label distributions are shown in Table 9.

A.2 News Stories (News)

The News dataset, introduced by Buechel et al. (2018), contains empathetic essay reactions to news articles with first-person empathy measurements collected by the following approach: First, the participants read a news article; second, they completed the Batson Empathic Concern – Personal Distress Scale (Batson et al., 1987), a multi-item scale for measuring first-person empathy, and continuous *empathy* and *distress* scores are derived empirically from this scale; and third, they wrote 300-800 character essays about the article. By associating the Batson empathy and distress scores with the text, the dataset offers a pairing of text-to-empathy scores of the actual text writer rather than annotated by a third-person observer. Given these essays, we perform the tasks of predicting the empathy and distress scores with regression models (News Empathy and News Distress). The label distributions are in Table 9.

Tafreshi et al. (2021) extended the News dataset to include person-level demographic and personality information and additional emotion labels. The emotion labels are based on Ekman's six basic emotions (*sadness, joy, disgust, surprise, anger,* and *fear*) (Ekman, 1971), with added *neutral* and *hope* labels. We perform a classification task (News Emotion) to predict these emotion labels for the essay. The News Emotion label distributions are in Table 10.

A.3 Empathic Conversations (Conv)

The Conv dataset (Omitaomu et al., 2022; Barriere et al., 2023) contains crowd-sourced conversations between two participants about articles in the News dataset described above. Before the conversation, each participant underwent the same process described for the News dataset. The conversations contain turn-level annotations of emotional polarity–(1) positive, (2) neutral, or (3) negative; emotional intensity–(1) no emotion, (2) weak emotion, (3) moderately strong emotion; (4) very strong emotion; or (5) extremely strong emotion, and empathy, as the degree to which the speaker puts themself in the shoes of the suffering victim–(1) not at all; (2) a little bit; (3) moderately; (4) quite a lot; and (5) extremely. Third-person observers of the conversation provided these annotations. Given the turn and prior dialogue context, we perform three separate regression tasks of predicting each of these assigned scores on each turn (Conv EmoPol, Conv EmoInt, and Conv Empathy). The label distributions for all Conv tasks are in Table 9.

A.4 Empathetic Dialogues (EmpDial)

The EmpDial dataset (Rashkin et al., 2019) contains dialogues based on emotional experiences and is widely used in NLP for fine-tuning generative empathy models. The speaker (target) first writes about a situation in which they experienced a given emotion. Then, they have a conversation with a listener (observer), in which they tell the story of the situation. The EmpDial conversations are accompanied by the grounding emotion and the situation written by the target in the public dataset. The dataset does not have empathy labels specifically, so we define two other turn-level tasks. First is role prediction (EmpDial Role), distinguishing between the speaker, who initiates the conversation based on a personal emotional situation, and the Listener, who responds without knowing the original emotion label or situation description. The second task is to predict the grounding emotion (EmpDial Emo). The model is provided the turn and prior dialogue context for both tasks. The label distributions for EmpDial Emo and EmpDial Role are in Table 12 and Table 10, respectively.

Other researchers extended the dataset with empathy labels; we use the **empathic response intents** (EmpDial EI) (Welivita and Pu, 2020) and the **empathic question intents** (EmpDial QI) (Svikhnushina et al., 2022) datasets described below, which are subsets of the original EmpDial dataset.

A.5 Empathic Intents (EmpDial EI)

EmpDial EI (Welivita and Pu, 2020) is a subset of EmpDial developed by Welivita and Pu (2020). It has nine different categorical labels of empathic intents: 1) questioning, 2) acknowledging, 3) neutral, 4) agreeing, 5) sympathizing, 6) encouraging, 7) suggesting, 8) consoling, and 9) wishing. While Welivita and Pu (Welivita and Pu, 2020) did not develop the scheme on a specific existing empathy construct, their work describes several existing theories from psychology and neuroscience which may inform the scheme, including Zillman's (2008) social emotion definition (Zillmann, 2008), *simulation theory* (Gordon, 1992), *theory-theory* (Gopnik and Wellman, 2012), and Singer and Klimecki's (2014) distinction between empathy and compassion (Singer and Klimecki, 2014). The nine intent labels are applied at the sentence level in turns of the EmpDial dataset. We perform the task of predicting the empathic intent labels at the sentence level, with the prior dialogue context provided as input. The EmpDial EI. label distributions are in Table 10.

A.6 Empathic Question Taxonomy (EmpDial QInt and EmpDial QAct)

After observing that *questioning* was the most frequent empathic intent among the prior subset of EmpDial, Svikhnushina et al. (2022) developed a fine-grained scheme that focuses on the role of questions in empathic conversations, with two different label sets. One set of labels categorizes nine question *acts* (QAct) resembling dialogue acts: 1) Ask about antecedent, 2) Suggest a solution, 3) Request information, 4) Ask about consequence, 5) Positive rhetoric, 6) Negative rhetoric, 7) Suggest a reason, 8) Ask for confirmation, and 9) Irony. The other set focuses on twelve categories of empathic intents underlying questions (QInt) in empathetic conversations: 1) Express concern, 2) Express interest, 3) Moralize speaker, 4) Sympathize, 5) Amplify joy, 6) Amplify excitement, 7) Support, 8) De-escalate, 9) Offer relief, 10) Amplify pride, 11) Motivate, and 12) Pass judgement. According to Svikhnushina et al. (2022), the coding scheme is informed by prior question classification schemes and reference to the principles of emotional regulation (Gross, 2013). Their motivation for fine-grained analyses of questions in empathic

interactions draws from other findings from psychology (e.g., Huang et al., 2017) and linguistics (e.g., Freed, 1994; Enfield et al., 2010) regarding their social role. On this dataset, we perform a task for each label set to predict the labels at the sentence level; the model is provided with the full turn of the sentence and the prior dialogue context. The EmpDial Empathic Intents and Question Intents datasets contain manual and automatic labels using models trained on the manual labels. We train the source adapters on the manually labeled samples in our study. The label distributions for each task are in Table 11.

A.7 Motivational Interviewing (MI)

The MI dataset (Welivita and Pu, 2022) contains Reddit interactions between support seekers and support providers curated from several mental health-related subreddits. The support provider turns are expert-annotated using the MITI coding scheme (Moyers et al., 2003), which are counselor behavior categories. Thus, the underlying construct of empathy is drawn from the foundational work on Motivational Interviewing, a counseling style often used for behavior change therapy (Miller and Rollnick, 2012). We perform two tasks on this data: a binary classification of MI-adherent and non-adherent behaviors (MI Adherent) and a multi-label classification of fine-grained behaviors (MI Behavior). MI Behavior has fourteen possible categorical labels applied to sentences of support-provider responses. The MI adherent and non-adherent classes are subsets of the fourteen behaviors, which align with the principles of motivational interviewing (adherent) or not (non-adherent). The MI adherent behaviors include 1) *affirm*, 2) *give information*, 3) *complex reflection*, 4) *support*, 5) *closed question*, 6) *emphasize autonomy*, 7) *simple reflection*, 8) *advise with permission* and 9) *open question*. The non-adherent behaviors include 10) *advise without permission*, 11) *self-disclosure*, 12) *direct*, 13) *warn*, and 14) *confront*. MI adherent behaviors relate to empathic behaviors and expectations (Moyers and Miller, 2013; Lord et al., 2015; Pérez-Rosas et al., 2017), especially in the *reflection* behaviors where counselors express their empathic understanding of what the client is saying (McCambridge et al., 2011; Pollak et al., 2011). The label distributions for MI Behavior and MI Adherent are in Table 12 and Table 10, respectively.

A.8 Epitome

The Epitome dataset (Sharma et al., 2020) encompasses dialogue exchanges between mental health support seekers and support providers from Reddit. Empathy is treated as three discernible forms of *expressed* empathy, including cognitive and emotional components (Cuff et al., 2016; Selman, 1980): 1) Emotional Reactions (ER), 2) Interpretations (IP), and 3) Explorations (EX), each providing a unique perspective on empathetic responses. As defined by the authors, ERs involve emotional expressions such as warmth, compassion, and concern; IPs communicate an inferred understanding of feelings and experiences; EXs represent efforts to improve understanding by investigating unstated feelings and experiences in the post. We perform the tasks of predicting the empathy level for each category on the supporter responses: no communication (0), weak communication (1), and strong communication (2). The model is provided with the support seeker context prior to the supporter's turn; note, however, a study by Lee et al. (2023) found that models tend to ignore the speaker context, which we do not analyze in this work. The label distributions for each task are in Table 10.

A.9 Empathy and Hope

Empathy Hope is a dataset of geopolitical Tweets from India and Pakistan (Yoo et al., 2021). Yoo et al.'s (2021) work focused on the health crisis emerging from the COVID-19 pandemic and the expressions of solidarity across country lines. They developed the dataset for detecting *supportive* content, which the dataset creators define as content expressing empathy, distress, or solidarity. This dataset's empathy construct is based on the News dataset's construct (Buechel et al., 2018), defining empathy as a "warm, tender, and compassionate feeling for a suffering entity," and distress as "a self-focused, negative affective state that occurs when one feels upset due to witnessing an entity's suffering or need." They apply a model developed on Buechel et al.'s (2018) data of empathy and distress, in addition to a hope speech classifier (Palakodety et al., 2020), which they use as signals for detecting annotated labels of supportive versus not-supportive content. We perform the task of predicting the manual labels of supportive versus not-supportive content. The label distributions are shown in Table 10.

Task	Definition		Link		Conducive	
	A	B	A	B	A	B
Condolence	5	5	2	2	3	4
Conv EmoInt	3	3	2	1	4	3
Conv EmoPol	3	3	2	2	4	5
Conv Empathy	5	5	4	3	5	5
EmpDial EI.	4	5	5	5	2	4
EmpDial Emo	2	2	1	1	1	3
EmpDial Role	1	2	1	1	1	2
Empathy Hope	2	3	3	2	1	1
Epitome ER	3	3	3	3	3	4
Epitome EX	3	3	3	3	3	4
Epitome IP	3	3	3	3	3	4
EmpDial QAct	3	3	5	3	2	4
EmpDial QInt	4	4	5	5	2	4
MI Adherent	2	3	3	2	3	3
MI Behavior	3	4	4	4	3	4
News Distress	4	5	4	4	4	4
News Emotion	3	3	3	2	4	3
News Empathy	5	5	5	4	4	4

Table 3: **Results from the construct aspect rating annotation task**. Here shows the 5-point Likert scale construct aspect ratings from each annotator A and B for each (Definition, Link, Conduciveness) on each task. Generally, the annotator ratings differ by 0 or 1 point; refer to Figure 4 for difference counts by annotator pair.

Metric	Definition		Link		Conduciveness		Overall	
Krippendorff's α	0.86		0.83		0.46		0.72	
	r	p-value	r	p-value	r	p-value	r	p-value
Spearman's r	0.91	0.00	0.89	0.00	0.49	0.04	0.70	0.00

Table 4: **Krippendorff's α values for measuring agreement between the annotators, and Spearman's rank correlation coefficients r between the Likert scores provided by the annotators.** The agreement is high on Definition and Link, but there is more disagreement on Conduciveness. The correlations are statistically significant (p-value ≤ 0.05), with positive coefficients across each aspect.

B Annotations and Construct Groups

Table 2 in the main text shows the separate themes and their corresponding annotations for definition, link, and conduciveness. Note, we find that the cutoffs for the means are close to linearly separating the three groups, with a clear cutoff between direct and abstract, but a less clear distinction between abstract and adjacent.

C Training Details

We ran a hyperparameter search using all empathy tasks; the best hyperparameter configurations for each empathy task are shown in Table 5. Models were run for ten epochs, with learning rates in the range of $1e^{-3}$ to $1e^{-6}$, batch size in the range of 8 to 64, taking the validation set performance from the best epoch. We considered the adapter model type as a hyperparameter. We tested the bottleneck adapter (Pfeiffer et al., 2020), which adds layers after the feed-forward block in each transformer layer; the double bottleneck (Houlsby et al., 2019), which adds layers both before and after the feed-forward block; prefix tuning (Li and Liang, 2021), which adds trainable prefix parameters to the keys and values in the attention heads; prompt tuning (Lester et al., 2021), which instead appends tunable tokens to the input text; compacter (Mahabadi et al., 2021), which adds parameterized hypercomplex multiplication layers (Zhang et al., 2021) in place of the feed-forward adapter layers; and LoRA (Hu et al., 2022), which instead uses

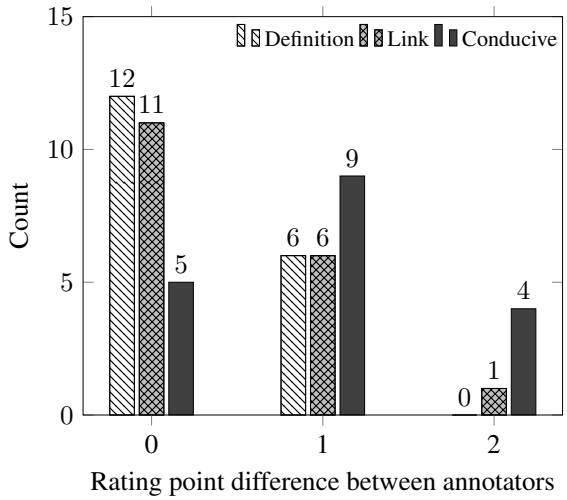

Figure 4: **Counts of differences in ratings by the annotators on each construct aspect.** For Definition and Link, the annotators most frequently differ by zero or one point. However, Conduciveness has a larger degree of disagreement, and the annotators more frequently differ by one point.

low-rank decomposition matrices in the attention layers. After 500 trials, we found that bottleneck adapters and LoRA outperformed other methods in all cases. We used the best batch size and learning rate for each task and trained these three adapters on all empathy tasks. The single bottleneck adapter performed best most of the time across task validation sets. Since we intend to stack adapters and compare task embeddings, we used this adapter type for all subsequent experiments.

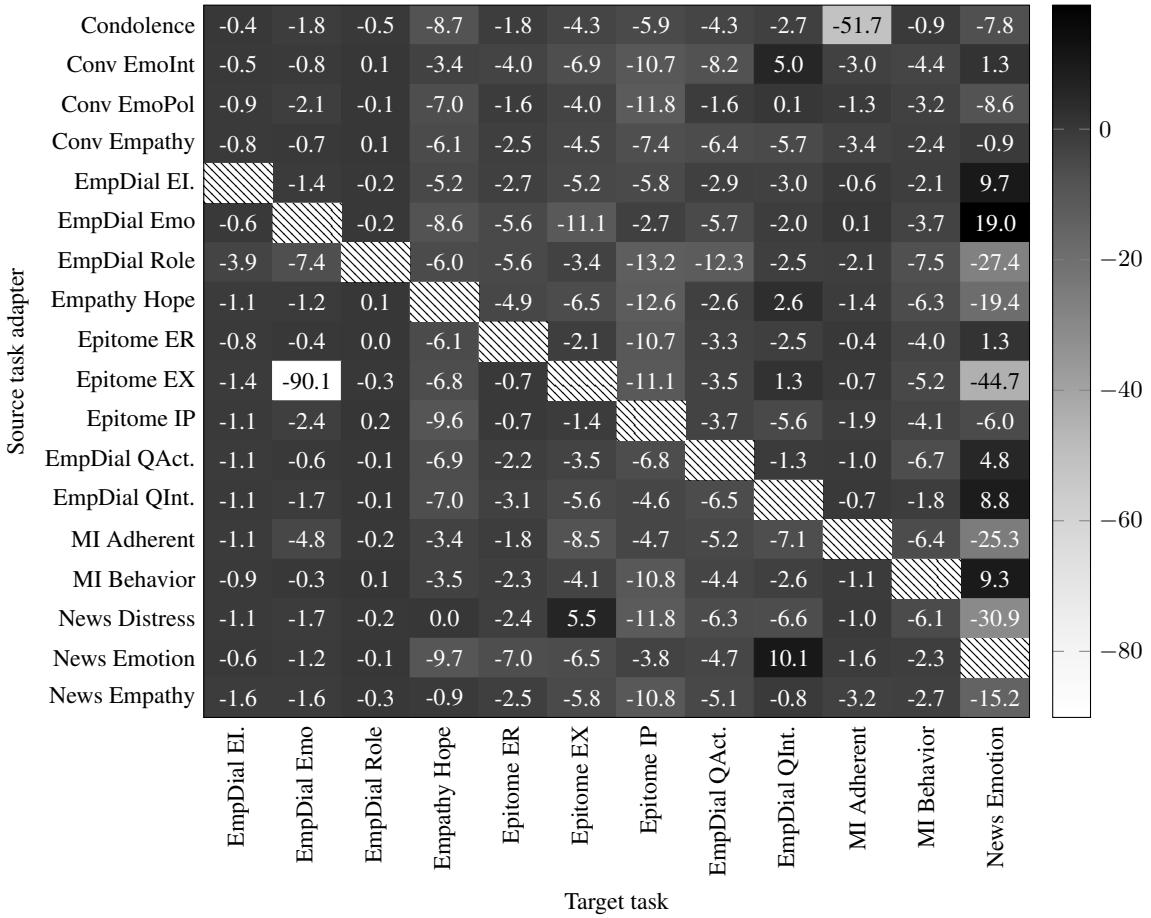

Figure 5: Classification task heat map of transfer performance of empathy source adapters on empathy target tasks. Values are calculated as the percent improvement of the transfer model over the baseline performance measured by the F1 Score.

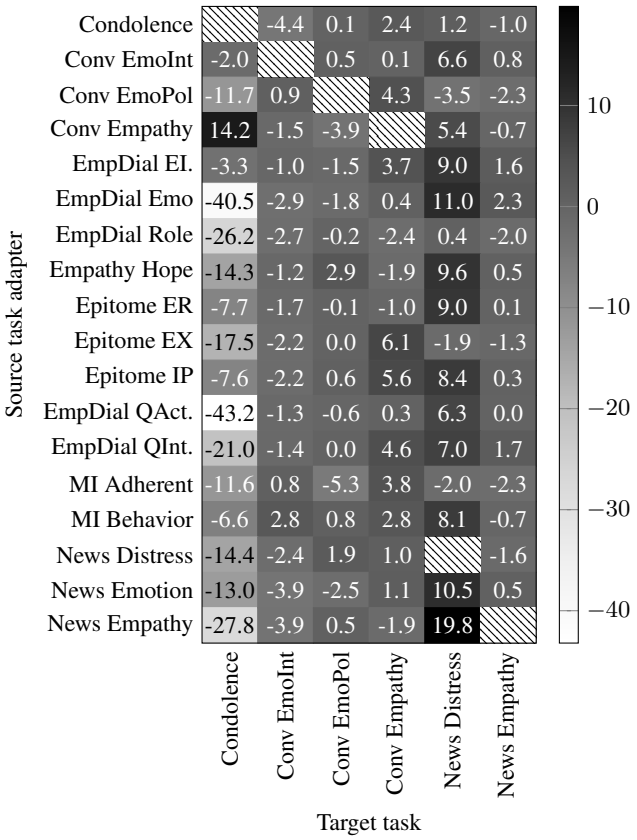

Figure 6: Regression task heat map of transfer performance of empathy source adapters on empathy target tasks. Values are calculated as the percent improvement of the transfer model over the baseline performance measured by Pearson r correlation.

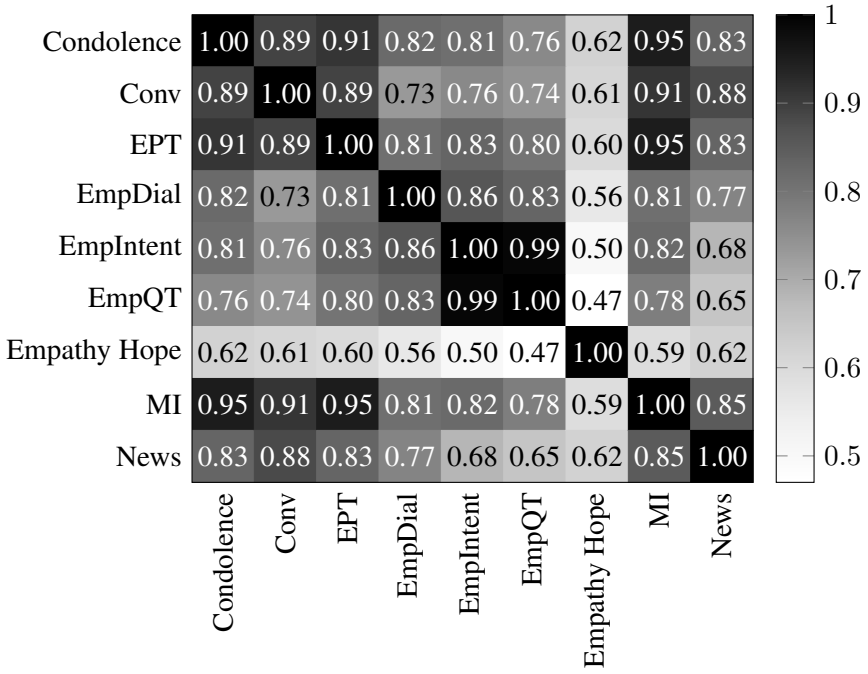

Figure 7: Heat map of dataset cosine similarity. Calculated with SBERT embeddings averaged over all dataset instances.

Task	Epoch	Batch Size	Learning Rate
Condolence	8	32	0.00028
Conv EmoInt	7	54	0.00083
Conv EmoPol	8	32	0.00093
Conv Empathy	8	31	0.00083
EmpDial EI.	11	44	0.00073
EmpDial Emo	9	18	0.00069
EmpDial Role	11	12	0.00086
Empathy Hope	7	47	0.00064
Epitome ER	7	42	0.00041
Epitome EX	10	34	0.00090
Epitome IP	9	22	0.00078
EmpDial QAct.	8	24	0.00056
EmpDial QInt.	10	52	0.00078
MI Adherent	8	10	0.00076
MI Behavior	10	30	0.00048
News Distress	6	30	0.00087
News Emotion	6	15	0.00043
News Empathy	11	14	0.00045

Table 5: Hyperparameter configurations for the best adapters for each empathy task on the validation sets.

Feature	Importance
C1 Definition	0.399 ± 0.095
C2 Link	0.292 ± 0.091
Data Source	0.228 ± 0.068
Language Setting	0.176 ± 0.067
C3 Conducive	0.105 ± 0.055
Number Labels	0.087 ± 0.042
Task Type	0.039 ± 0.025
Sample Size	0.028 ± 0.021
Token Size	0.026 ± 0.027
Vocab Size	0.023 ± 0.021
Data Emb Sim	0.001 ± 0.001
Task Emb Sim	0.000 ± 0.000

Table 6: Feature importances in regression fit to improvement over baseline calculated as the difference in R^2 when permuting the feature.

	EmpDial			Empathy	Epitome			EmpDial		MI		News
	EI	Emo	Role	Hope	ER	EX	IP	QAct	QInt	Adherent	Behavior	Emotion
Condolence	.88	.38	.94	.84	.81	.76	.59	.58	.38	.40	.53	.67
Conv EmoInt	.88	.38	.95	.89	.79	.74	.56	.56	.41	.81	.51	.74
Conv EmoPol	.88	.38	.94	.86	.81	.76	.56	.60	.39	.82	.51	.66
Conv Empathy	.88	.38	.95	.87	.80	.76	.59	.57	.36	.81	.52	.72
EmpDial EI.	**.89**	.38	.94	.87	.80	.75	.60	.59	.37	.83	.52	.80
EmpDial Emo	.88	**.39**	.94	.84	.78	.70	.61	.57	.38	**.83**	.51	**.87**
EmpDial Role	.85	.36	.95	.87	.78	.77	.55	.53	.38	.82	.49	.53
Empathy Hope	.88	.38	.95	**.92**	.78	.74	.55	.59	.40	.82	.50	.59
Epitome ER	.88	.38	.95	.87	**.82**	.77	.56	.59	.38	.83	.51	.74
Epitome EX	.87	.04	.94	.86	.82	.79	.56	.59	.39	.83	.50	.40
Epitome IP	.88	.38	**.95**	.83	.82	.78	**.63**	.58	.36	.82	.51	.68
QAct. Manual	.88	.38	.94	.86	.81	.76	.59	**.61**	.38	.83	.50	.76
QInt. Manual	.88	.38	.94	.86	.80	.75	.60	.57	.39	.83	.52	.79
MI Adherent	.88	.37	.94	.89	.81	.72	.60	.58	.36	.83	.50	.54
MI Behavior	.88	.38	.95	.89	.80	.76	.56	.58	.38	.82	**.53**	.79
News Dis	.88	.38	.94	**.92**	.80	**.84**	.56	.57	.36	.83	.50	.50
News Emotion	.88	.38	.95	.83	.77	.74	.61	.58	**.43**	.82	.52	.73
News Empathy	.87	.38	.94	.91	.80	.75	.56	.58	.38	.81	.52	.62

Table 7: F1 scores for each classification target task (columns). The baseline scores (i.e., source is the same as target) are underlined, and the best result is in bold.

	Condolence	Conv EmoInt	Conv EmoPol	Conv Empathy	News Dis	News Empathy
Condolence	.43	.71	.73	.65	.72	.91
Conv EmoInt	.42	.74	.73	.64	.75	.93
Conv EmoPol	.38	.74	.73	.67	.68	.90
Conv Empathy	**.49**	.73	.70	.64	.75	.91
EmpDial EI.	.42	.73	.72	.66	.77	.93
EmpDial Emo	.26	.72	.72	.64	.79	**.94**
EmpDial Role	.32	.72	.73	.62	.71	.90
Empathy Hope	.37	.73	**.75**	.63	.78	.92
Epitome ER	.40	.73	.73	.63	.77	.92
Epitome EX	.36	.72	.73	**.68**	.69	.91
Epitome IP	.40	.72	.73	.67	.77	.92
QAct. Manual	.25	.73	.73	.64	.75	.92
QInt. Manual	.34	.73	.73	.67	.76	.94
MI Adherent	.38	.74	.69	.66	.69	.90
MI Behavior	.40	**.76**	.73	.66	.77	.91
News Dis	.37	.72	.74	.64	.71	.90
News Emotion	.38	.71	.71	.65	.78	.92
News Empathy	.31	.71	.73	.63	**.85**	.92

Table 8: Pearsonr scores for each regression target task (columns). The baseline scores (i.e., source is the same as target) are underlined, and the best result is in bold.

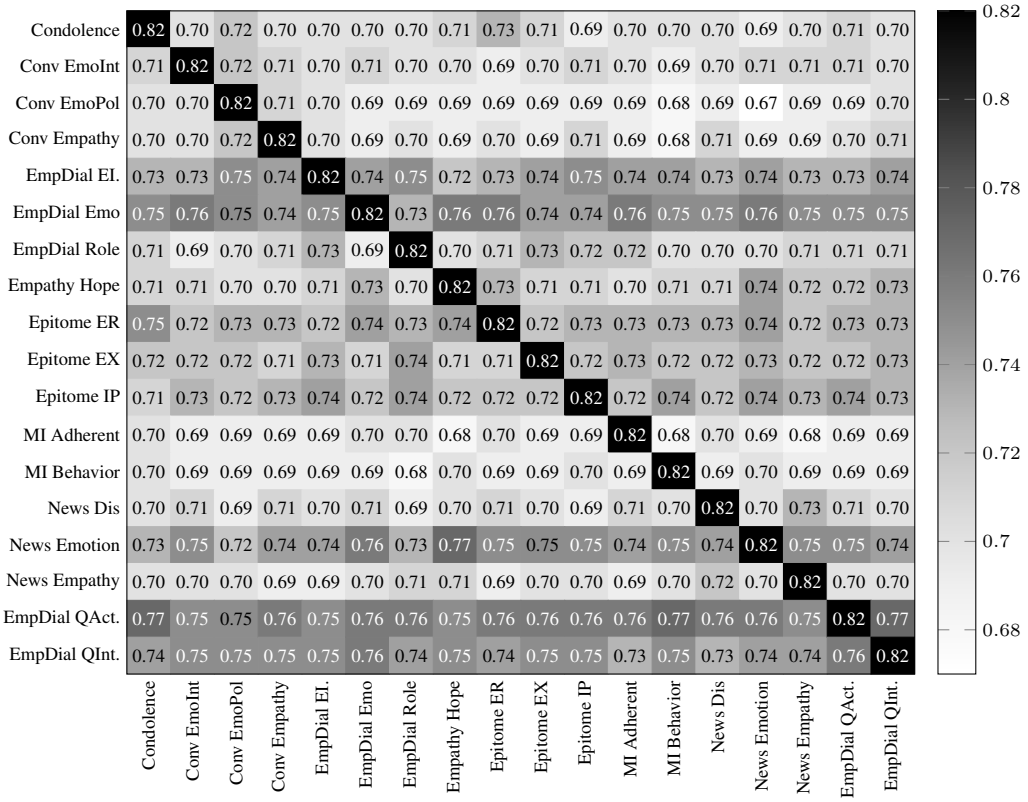

Figure 8: Heatmap of the task embedding similarity rankings. These are computed by ranking the cosine similarities between specific model layers of each task and then computing RRF over those rankings.

Instructions for Empathy Operationalization Annotation Task

CONTEXT

We are investigating how empathy constructs are operationalized for NLP research via measurable or observable characteristics in language. We aim to understand whether the various operationalizations of the construct are convergent and valid, capturing aspects of the targeted construct. In order to characterize the operationalizations, we are collecting human annotated ratings for 1) the level of granularity of the construction definition, 2) the correspondence of the measurements or observations in language to the construct definition, and 3) the Conduciveness of the language scenario to represent empathic processes in the language data. The annotation task requires understanding these sources' scientific methods and experimental design, namely, how each paper described the construct of empathy, their empathy measurement or observation methods, and the experimental setting from which the data was collected or obtained. Thus, background knowledge of empathy constructs in psychology and the landscape of empathy research in NLP is necessary; prior knowledge of each dataset and task is beneficial.

INSTRUCTIONS

We will provide descriptions of 18 empathy language tasks from 9 different datasets and references to the original articles where they were proposed. Note that some tasks are defined here and not in the original paper, which we added to support our study. You will be asked to rate these three aspects on a 5-point Likert scale:

Each task/dataset will be annotated according to three aspects on a 5-point Likert Scale:

1. Definition: How granular (coarse/high level to fine/more detailed) is the empathy construct or theory that grounds the data development or task definition?
2. Link/Correspondence: To what degree do the measurement and/or observation approaches correspond to the construct's defined components?
3. Conduciveness: How does the language scenario/communication context* influence your expectation of observing the empathy construct in the data, considering aspects such as data collection methods (survey, interview, observation, experiment, etc.), annotation criteria, and social norms of the context?

**Language scenario* refers to properties of the corpus of communications, such as: What form of communication does the data reflect? Is the corpus a collection of conversations–if so, are they asynchronous, goal-oriented, etc? Is it a collection of interactions in online forums–if so, what purpose do the online forums serve, or are there various purposes? Is it a collection of messages in isolation from an interaction (e.g., an essay)? What was the data collection method (e.g., experiment, survey, observation)? Does the corpus center on a particular domain?

Next, we will provide the scoring criteria for rating these aspects on the 5-point Likert scale (see Figure 10).

Figure 9: Instructions for the empathy operationalization annotation task.

Scoring Criteria for Empathy Operationalization Categorization

Definition.

1. The empathy concept is neither defined nor described nor is there a referenced source from which a definition is drawn.
2. The empathy concept defined or referenced bypasses its multidimensionality, e.g., by focusing on a single aspect without relating it to or discussing other possible aspects, by being abstract, simplified, or vague.
3. The empathy concept defined or referenced acknowledges more than one dimension of empathy but leaves these dimensions abstract, e.g., by merely referring to emotional and cognitive empathy.
4. The empathy concept has a fairly fine granularity. It comprises more than one dimension of empathy, such as cognitive and emotional empathy, with high-level descriptions of how they emerge in language or are experienced and perceived.
5. The empathy concept has a very fine granularity in the dimensions/factors/aspects hypothesized to comprise empathic experiences or interactions by including several concrete dimensions of empathic experiences, behaviors, or interactions.

Link/Correspondence.

1. The measurements or observations have a *very weak* correspondence to the defined construct. They may capture aspects of empathy or related phenomena, but not of the defined empathy construct by any direct or indirect approach.
2. The measurements or observations have a *weak* correspondence to the defined construct. They seem intended to reflect the empathy construct, but the connection is unclear, or a high level of abstraction likely interferes with capturing what is intended.
3. The measurements or observations *somewhat* correspond to the defined construct. They clearly reflect the defined construct, but there is a fair amount of abstraction; some elements of the construct are missing, or the methods may not be reliable.
4. The measurements or observations correspond *fairly well* to the defined construct. There may be some abstraction, but they are systematically derived based on several aspects of the construct with moderately reliable methods.
5. The measurements or observations correspond *very well* to the defined construct; they directly correspond to the individual facets of the construct, and the methods seem highly appropriate and precise for measuring what is intended.

Conduciveness.

1. Aspects of the scenario *significantly* lower my expectations of observing empathy.
2. Aspects of the scenario *somewhat* lower my expectations of observing empathy.
3. The properties of the scenario do not affect my expectations of observing empathy.
4. Aspects of the scenario *somewhat* increase my expectations of observing empathy.
5. Aspects of the scenario *significantly* increase my expectations of observing empathy.

Figure 10: Criteria provided to the annotators for scoring each aspect of the empathy construct along a 5-point Likert scale.

	Condolence				**Conv EmoInt**				
	All	Train	Val	Test		All	Train	Val	Test
Split	100.0	79.8	10.2	10.1	Split	100.0	80.1	10.3	9.6
Min	1.00	1.00	1.00	1.00	Min	0.00	0.00	0.00	0.00
Median	1.50	1.50	1.50	1.50	Median	2.00	2.33	2.00	2.00
Max	5.00	5.00	5.00	5.00	Max	5.00	5.00	5.00	4.67
Mean	1.54	1.53	1.57	1.56	Mean	2.22	2.24	2.19	2.09
std	0.76	0.75	0.83	0.82	std	0.82	0.82	0.88	0.79
	Conv EmoPol				**Conv Empathy**				
	All	Train	Val	Test		All	Train	Val	Test
Split	100.0	80.1	10.3	9.6	Split	100.0	80.1	10.3	9.6
Min	0.00	0.00	0.00	0.00	Min	0.00	0.00	0.00	0.00
Median	1.33	1.33	1.33	1.33	Median	2.00	2.00	2.00	2.00
Max	2.00	2.00	2.00	2.00	Max	5.00	5.00	4.67	4.33
Mean	1.29	1.29	1.28	1.28	Mean	2.09	2.08	2.17	2.02
std	0.61	0.61	0.61	0.60	std	0.92	0.92	0.92	0.91
	News Distress				**News Empathy**				
	All	Train	Val	Test		All	Train	Val	Test
Split	100.0	29.6	37.0	33.3	Split	100.0	29.6	37.0	33.3
Min	1.00	1.00	1.00	1.00	Min	1.00	1.00	1.00	1.00
Median	3.75	3.62	3.75	3.69	Median	4.33	4.33	4.33	4.33
Max	7.00	7.00	7.00	7.00	Max	7.00	7.00	7.00	7.00
Mean	3.76	3.74	3.79	3.74	Mean	4.26	4.25	4.26	4.25
std	2.00	2.01	2.00	2.00	std	1.95	1.94	1.95	1.95

Table 9: Measures of central tendency of the rating values in the regression tasks, overall and per experimental split.

EmpDial EI.

	All	Train	Val	Test
Split	100.0	80.0	10.0	10.0
questioning	32.5	32.7	31.7	32.3
acknowledging	20.6	20.5	20.9	20.8
neutral	14.7	14.8	14.4	14.4
agreeing	13.0	13.0	12.9	13.2
sympathizing	4.6	4.6	4.8	4.4
encouraging	4.1	4.0	4.5	4.1
suggesting	3.9	3.9	4.0	4.2
consoling	3.7	3.7	3.8	3.9
wishing	2.9	2.9	3.0	2.7

EmpDial Role

	All	Train	Val	Test
Split	100.0	80.1	9.9	10.0
Speaker	62.2	62.2	62.5	62.2
Listener	37.8	37.8	37.5	37.8

Epitome ER

	All	Train	Val	Test
Split	100.0	80.0	10.0	10.0
0	66.4	66.3	67.0	66.6
1	28.6	28.8	27.7	28.5
2	5.0	4.9	5.3	5.0

Epitome IP

	All	Train	Val	Test
Split	100.0	80.0	10.0	10.0
0	52.5	52.7	51.5	52.0
1	3.8	3.8	3.6	3.6
2	43.7	43.5	44.9	44.4

News Emotion

	All	Train	Val	Test
Split	100.0	29.6	37.0	33.3
Sadness	38.3	39.1	37.5	38.4
Neutral	27.2	26.8	27.3	27.4
Anger	16.3	16.8	16.2	16.0
Disgust	8.4	7.9	8.8	8.4
Hope	4.2	4.0	4.4	4.1
Fear	3.8	3.8	3.8	3.8
Joy	1.1	1.1	1.1	1.1
Surprise	0.7	0.6	0.9	0.7

MI Adherent

	All	Train	Val	Test
Split	100.0	79.7	10.2	10.2
Adherent	68.5	68.7	67.9	67.5
Non-adherent	31.5	31.3	32.1	32.5

Epitome EX

	All	Train	Val	Test
Split	100.0	80.0	10.0	10.0
0	84.4	84.4	84.5	84.1
1	3.4	3.4	3.3	3.3
2	12.2	12.2	12.2	12.6

Empathy and Hope

	All	Train	Val	Test
Split	100.0	80.0	10.1	10.0
Supportive	50.7	50.7	50.4	50.8
Not Supportive	49.3	49.3	49.6	49.2

Table 10: Label distributions overall and per each experimental split for the EmpDial EI, EmpDial Role, News Emotion, MI Adherent, Epitome, and Empathy and Hope tasks.

EmpDial QAct

	All	Train	Val	Test
Split	100.0	80.0	10.0	10.0
Request information	51.4	51.5	50.3	51.3
Ask about consequence	17.9	18.1	17.3	17.6
Ask about antecedent	11.3	11.1	12.3	12.1
Suggest a solution	8.0	8.0	8.2	8.0
Ask for confirmation	5.2	5.2	4.9	5.1
Suggest a reason	4.1	4.1	4.7	3.9
Positive rhetoric	1.1	1.1	1.1	1.0
Negative rhetoric	0.8	0.7	1.0	0.7
Irony	0.2	0.2	0.1	0.1

EmpDial QInt

	All	Train	Val	Test
Split	100.0	80.0	10.0	10.0
Express interest	60.2	60.4	59.2	59.8
Express concern	23.4	23.2	23.7	23.9
Sympathize	5.1	5.0	5.9	4.9
Offer relief	4.5	4.6	4.1	4.7
Amplify excitement	2.3	2.3	2.7	2.2
Support	1.0	1.0	1.0	1.0
Amplify joy	0.9	0.8	0.9	1.0
Amplify pride	0.7	0.7	0.7	0.7
De-escalate	0.7	0.7	0.7	0.6
Moralize speaker	0.6	0.6	0.5	0.5
Pass judgement	0.5	0.5	0.4	0.5
Motivate	0.2	0.2	0.1	0.1

Table 11: Label distributions overall and per each experimental split for the EmpDial QAct and QInt tasks.

EmpDial Emotion	All	Train	Val	Test
Split	100.0	80.1	9.9	10.0
surprised	5.2	5.1	5.0	5.7
excited	3.8	3.8	3.3	3.6
annoyed	3.6	3.6	3.4	3.7
proud	3.5	3.5	3.1	4.1
angry	3.5	3.5	2.9	3.9
sad	3.5	3.4	3.9	3.4
lonely	3.3	3.2	3.5	3.2
grateful	3.2	3.2	3.4	3.2
afraid	3.2	3.2	3.1	3.4
confident	3.2	3.2	3.2	3.4
disgusted	3.2	3.2	3.2	3.2
impressed	3.2	3.2	2.8	3.2
terrified	3.2	3.2	3.2	2.6
anxious	3.1	3.2	3.0	2.9
disappointed	3.1	3.1	3.5	3.0
anticipating	3.1	3.0	3.2	3.9
hopeful	3.1	3.1	3.4	3.0
jealous	3.1	3.1	3.4	2.8
guilty	3.1	3.0	3.3	3.3
joyful	3.1	3.1	3.1	3.0
furious	3.0	3.0	2.8	3.4
nostalgic	3.0	3.0	3.2	2.9
prepared	3.0	3.1	2.9	2.9
embarrassed	3.0	3.0	2.8	2.7
content	2.9	3.0	2.9	2.5
sentimental	2.8	2.9	3.0	2.6
devastated	2.8	2.9	2.5	2.7
caring	2.7	2.7	3.0	2.9
trusting	2.6	2.7	2.3	2.7
ashamed	2.6	2.5	3.3	2.3
apprehensive	2.5	2.5	2.8	2.2
faithful	1.9	1.9	1.7	1.8

MI Behavior	All	Train	Val	Test
Split	100.0	79.7	10.2	10.2
Give Information	29.9	30.4	28.0	27.9
Advise w/o Permission	14.3	14.0	16.0	14.9
Self-Disclose	8.8	8.8	7.9	9.4
Complex Reflection	8.0	7.8	9.6	8.3
Support	7.8	7.9	7.5	7.7
Affirm	5.9	5.9	5.1	6.7
Closed Question	5.8	5.8	5.7	5.3
Direct	5.7	5.8	5.0	5.7
Simple Reflection	3.5	3.4	4.1	3.4
Open Question	3.0	3.0	3.3	3.2
Advise w/ Permission	3.0	2.9	3.3	3.4
Confront	2.0	2.0	2.4	1.8
Emphasize Autonomy	1.6	1.6	1.4	1.6
Warn	0.7	0.7	0.8	0.8

Table 12: Label distributions overall and per each experimental split for the EmpDial Emotion and MI Behavior tasks.