LeadEmpathy: An Expert Annotated German Dataset of Empathy in Written Leadership Communication

Didem Sedefoglu[1], **Allison Lahnala**[2], **Jasmin Wagner**[1]
Lucie Flek[2], **Sandra Ohly**[1]
[1]Department of Business Psychology, University of Kassel, Germany
[2]Conversational AI and Social Analytics (CAISA) Lab, University of Bonn, Germany
didem.sedefoglu@uni-kassel.de, alahnala@uni-bonn.de

Abstract

Empathetic leadership communication plays a pivotal role in modern workplaces as it is associated with a wide range of positive individual and organizational outcomes. This paper introduces LeadEmpathy, an innovative expert-annotated German dataset for modeling empathy in written leadership communication. It features a novel theory-based coding scheme to model cognitive and affective empathy in asynchronous communication. The final dataset comprises 770 annotated emails from 385 participants who were allowed to rewrite their emails after receiving recommendations for increasing empathy in an online experiment. Two independent annotators achieved substantial inter-annotator agreement of $\geq$ .79 for all categories, indicating that the annotation scheme can be applied to produce high-quality, multidimensional empathy ratings in current and future applications. Beyond outlining the dataset's development procedures, we present a case study on automatic empathy detection, establishing baseline models for predicting empathy scores in a range of ten possible scores that achieve a Pearson correlation of 0.816 and a mean squared error of 0.883. Our dataset is available at `https://github.com/caisa-lab/LEAD-empathy-dataset`.

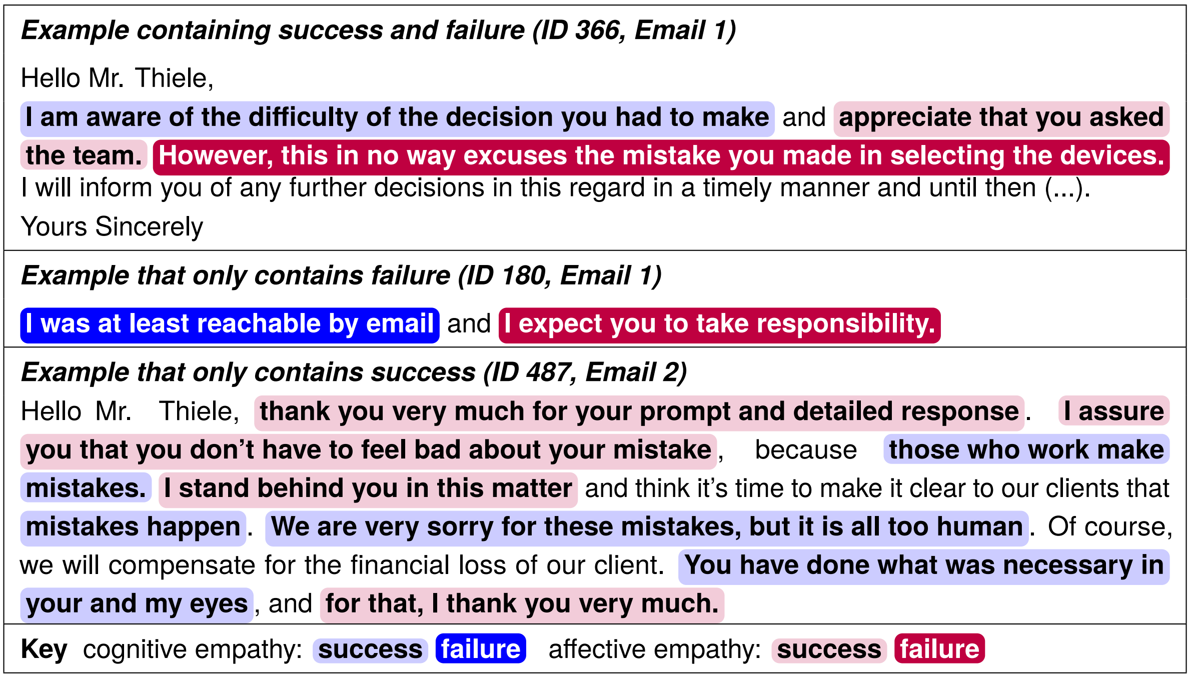
Keywords: affective empathy, cognitive empathy, empathy annotation scheme, leadership communication

1. Introduction

In the dynamic and increasingly digitized organizational landscape, the communication between leaders and followers has been increasingly shifting to virtual contexts (Bell et al., 2023). To enable meaningful interactions at work, leaders need to communicate effectively using digital technologies (Tuschner et al., 2022). Given that nonverbal cues such as body language and mimics are mostly absent in computer-mediated communication, this presents a notable challenge (Marlow et al., 2017). In light of this, the ability to convey empathy through written communication emerges as an important facet of effective leadership. Empathy, described as the ability to recognize and understand another person's feelings and share corresponding emotions (Cuff et al., 2016), enhances the quality of relationships between leaders and followers (Mahsud et al., 2010), and is positively associated with follower satisfaction and performance (Kock et al., 2019). Employee perceptions of leaders' empathy is determined by the behavior that is demonstrated by leaders, with a large portion manifesting through empathetic communication (Clark et al., 2018). Hence, organizations should prioritize the development and training of their leaders' empathetic communication skills.

Leveraging language-based models offers a promising approach to enhancing leaders' empathetic communication skills in virtual contexts. AI-driven tools have the benefit of being accessible at any time, providing personalized real-time feedback and immediate recommendations that leaders can use to self-reflect and adjust their behavior (Quaquebeke and Gerpott, 2023). Evidence from the mental health and education domain underscores that timely feedback from an AI can be an effective tool for developing empathetic communication skills. For example, prior research has shown that human-AI collaboration can help achieve higher levels of empathy in text-based peer-to-peer mental health support (Sharma et al., 2020) and student peer reviews (Wambsganss et al., 2021).

Effective AI-driven feedback tools for empathetic communication heavily rely on robust natural language processing (NLP) models, necessitating significant contributions from the NLP field. Firstly, empathy detection models are essential to evaluate the level of empathy and provide insights into the linguistic and conversational behaviors underlying successful empathic expression (Pérez-Rosas et al., 2017; Sharma et al., 2020; Wu et al., 2021; Wambsganss et al., 2022). Secondly, integrating generative models can provide specific recommendations for enhancing empathy in a given expression or passage of text (Sharma et al., 2021), such as leadership emails. Training these models necessitates informed frameworks of empathy and reliable measurement approaches for constructing quality language resources (Lahnala et al., 2022), which given the subjectivity of perceived empathy remains an open challenge. However, the need



Example containing success and failure (ID 366, Email 1)

Hello Mr. Thiele,
I am aware of the difficulty of the decision you had to make and **appreciate that you asked the team.** **However, this in no way excuses the mistake you made in selecting the devices.**
I will inform you of any further decisions in this regard in a timely manner and until then (...).
Yours Sincerely

Example that only contains failure (ID 180, Email 1)

I was at least reachable by email and **I expect you to take responsibility.**

Example that only contains success (ID 487, Email 2)

Hello Mr. Thiele, **thank you very much for your prompt and detailed response**. **I assure you that you don't have to feel bad about your mistake**, because **those who work make mistakes.** **I stand behind you in this matter** and think it's time to make it clear to our clients that **mistakes happen**. **We are very sorry for these mistakes, but it is all too human**. Of course, we will compensate for the financial loss of our client. **You have done what was necessary in your and my eyes**, and **for that, I thank you very much.**

Key cognitive empathy: **success** **failure** affective empathy: **success** **failure**

Figure 1: Translated examples from the dataset demonstrating the application of the coding scheme. The examples in the original German are shown in Figure 4 in the Appendix.

for comprehensive, fine-grained empathy datasets encompassing diverse contexts and distinguishing nuanced constructs such as cognitive and affective empathy is a significant shortcoming in the current NLP research landscape. Addressing this limitation is crucial for advancing empathetic language models that can be integrated with AI-driven feedback tools.

Furthermore, computational understanding of empathetic communication calls for investigations of empathy across various communicative domains with variations of social properties and communicative intents that span different languages and cultures. While there is an increasing body of work concerning NLP models for empathy that explore various communicative domains, primarily focusing on English-language open-domain dialogues and clinical or mental health settings, empathy in written leadership communication remains untouched.

Contributions. We summarize our contributions and the value of LeadEmpathy for NLP research as follows:

1. An innovative data collection procedure and annotation scheme with specific behavioral indicators based on psychological theory and preexisting scales for empathy detection (Sharma et al., 2021; Amjad et al., 2023).
2. A novel empathy dataset that i) has paired examples (email1 & email2)[1] and ii) is in German contributing to the gap a empathy resources for non-English languages. and iii) to our knowledge, is the first available dataset for empathy in written leadership communication. Translated examples are shown in Figure 1.
3. A case study demonstrating its utility for the development of computational empathy models and technology via an empathy detection task.

2. Related Work

2.1. Empathy conceptualization

Empathy, a central construct in various fields including psychology, neuroscience, sociology, and medical research, has long suffered from a lack of consensus regarding its definition and measures. To illustrate the variation in definitions, Cuff et al. (2016) conducted a comprehensive review, identifying 43 distinct definitions of empathy. Today, most scholars recognize empathy as a multidimensional construct containing both affective and cognitive elements, while some definitions also include a behavioral dimension (Clark et al., 2018).

In describing the dimensions of empathy, in line with Davis (1983) terminology, we differ between (i) *target* as the person who is experiencing an emotion in a given situation, acting as the source for another person's empathy and (ii) *observer* as the person who perceives the target's affective state and is at the disposition for an empathetic experience.

[1] To our knowledge, there are no existing datasets annotated with empathy with pairs before and after empathy instructional interventions.

Cognitive empathy refers to the ability to understand another person's situation, thoughts, and emotional state (Clark et al., 2018). It involves strategies such as perspective-taking and retrieving relevant memories to gain insight into the mental states of others (Cuff et al., 2016). Affective empathy is the ability to experience similar emotions in response to observing another person's state (Clark et al., 2018). It involves processes such as emotional contagion and recognition through which affective states are transmitted from the target to the observer (Shamay-Tsoory, 2011). Behavioral empathy is the demonstration of cognitive empathy and affective empathy in both verbal and nonverbal behavior (Clark et al., 2018). Compared to affective and cognitive empathy, research on behavioral empathy is sparse. Despite the importance of empathy for effective social interactions at work (Mahsud et al., 2010; Kock et al., 2019), to date, it is unclear how affective and cognitive empathy manifest in behaviors, especially in the organizational context. Scales that have been previously developed are mostly designed for face-to-face interactions (Bylund and Makoul, 2005) and are context-specific, e.g. developed for online mental health platforms (Sharma et al., 2020) or criminal suspect interviews (Dando et al., 2016). Scales designed for in-person interactions contain a wider range of non-verbal empathetic behaviors such as nodding or using expressive voice. In comparison, showing empathy in text-based exchanges such as email or chat is limited to the content of the texts. In addition, changing the context might change the behaviors that are considered important. For example, Dando et al. (2016) investigated interviews conducted by police officers and suggested that in this specific context, empathy can be expressed by *proving spontaneous comfort*, which includes *offering a drink* or *additional time to answer*. While the category is legitimate in the respective context, it reinforces the point that existing scales need to be adapted to suit the type of communication assessed in our context.

Thus, we focus on developing and applying a scale for empathy expressed in text-based asynchronous leadership communication. In doing so, we follow Clark et al. (2018)'s call for new measures that capture valid observer ratings of empathetic communication in organizations.

2.2. Computational empathy

Research on empathetic language in NLP focuses on two broad tasks, empathy recognition and generation. Detection includes predicting empathy scores by regression models (Buechel et al., 2018), classifying the degree or presence empathy (e.g. low to high, empathy versus no empathy) (Sharma et al., 2020; Hosseini and Caragea, 2021), and predicting empathy-related labels (Welivita and Pu, 2020; Svikhnushina et al., 2022). For example, Buechel et al. (2018)'s dataset of reactions to news articles has scores for empathic concern and personal distress, and has been utilized for recent shared tasks on empathy detection (Tafreshi et al., 2021; Barriere et al., 2022).

Research on generative models has been explored for open-domain dialogue settings (Rashkin et al., 2019; Lin et al., 2019; Smith et al., 2020; Majumder et al., 2020; Naous et al., 2021; Welivita et al., 2021), customer care agents (Firdaus et al., 2020), and counseling (Shen et al., 2020), among others. Zhong et al. (2020), for example, developed a dataset and models for persona-based empathetic conversations across a variety of domains. Rather than generating full empathetic responses, Sharma et al. (2021) developed a model for empathic rewriting which makes sentence-level edits to a given expression. Such empathic rewriting models can be integrated into AI-driven tools for training empathic communication and providing feedback.

Challenges in computational research on empathetic language arise from how empathy is defined and measured. As much of this research is trained on large-scale datasets with abstract labels based on broad conceptualizations of empathy (Rashkin et al., 2019), some recent works aimed to integrate theory-grounded multidimensional aspects and measurement approaches in new language resources. Examples include Sharma et al. (2020)'s EPITOME coding scheme applied to mental health support conversations, Xie and Pu (2021)'s and Welivita and Pu (2020)'s large-scale dialogue dataset integrating labels for emotion regulation and empathic intents, and Svikhnushina et al. (2022)'s empathetic question taxonomy applied to the EmpatheticDialogues dataset (Rashkin et al., 2019). A particularly relevant example is Wambsganss et al. (2021)'s work which introduced a coding scheme for emotional and cognitive empathy and applied it to German peer reviews. They trained predictive models which they integrated into an adaptive writing support system that provides feedback.

The LeadEmpathy dataset brings this body of research to a novel domain of written leadership communication with a new approach to measuring empathy. Furthermore, our approach accounts not only for the presence and absence of empathy, but we also define and label *empathic failures*, which to the best of our knowledge has not been explored in NLP research.

3. Annotation

To explore empathy in text-based leadership communication, we develop a new annotation scheme

	Points	**01.**	**cognitive empathy**	**02.**	**affective empathy**
a. failure	-1	01.a.	cognitive empathy failure	02.a.	affective empathy failure
b. absence	0	b.	lack of empathy (symbolic category)		
c. success	+1	01.c.	cognitive empathy success	02.c.	affective empathy success

Table 1: A simplified overview of the final empathy coding scheme, excluding the working definitions, category definitions, indicators, and examples.

based on prior existing scales and theories of empathy and empathetic communication. In a collaborative effort of organizational psychology and computer science researchers, the primary aim was to create a scale that a) can be applied to the assessment of written communication, b) gives quantifiable results similar to an empathy score c) differentiates between specific behaviors associated with affective and cognitive empathy and d) is applicable to the the organizational context. The scale was developed using a deductive approach guided by procedural standards from qualitative content analysis (Mayring, 2014). Based on this approach, the level of empathy can be reliably assessed if all relevant markers, in our case verbal expressions indicating empathy, are extracted from literature and reflected in the coding scheme.

As a first step, we determined the broad structure of our scheme based on the definition of empathy (Clark et al., 2018). Thus, in our coding scheme, we differentiate between expressions indicative of cognitive empathy and expressions indicative of affective empathy.

One of the crucial considerations while developing the scheme was whether there was also a negative polarity of empathy. Prior literature provided support for the idea that there might be expressions that signal the opposite of empathy, i.e. denial or disconfirmation of the target's perspective (Bylund and Makoul, 2002). The rationale behind this can be explained using the example of friendliness. To come across as friendly in a conversation, it is advisable to use friendly expressions, but also to refrain from using unfriendly expressions. Analogously, we conclude that the level of empathy of a message increases with the use of expressions demonstrating empathy (here: *empathy success*), but decreases with the use of expressions that demonstrate the opposite (here: *empathy failure*).

Thus, our final coding scheme encompasses four main categories: *cognitive empathy success, affective empathy success, cognitive empathy failure*, and *affective empathy failure*, as well as a symbolic category for texts that don't have any type of empathetic expression (i.e. only factual information), which we call the *absence of empathy*. In line with this, each *success* or *failure* results in the addition or deduction of one point to the total score. The absence category serves as a symbolic category in the system that is not associated with the addition or subtraction of any points. Table **??** shows an overview of the scheme and the full scheme is shown in Table 5 in the Appendix0. In the following, we describe the categories in detail.

Cognitive empathy success. The category *cognitive empathy success* is defined as detecting, recognizing, and understanding others' cognitive and emotional states, meaning their thoughts, motifs, and feelings, with an emphasis on the observer's act of taking in the target's mental state.

In terms of behavioral indicators, this category includes emotion cognition and perspective-taking (Cuff et al., 2016), understanding (Watzlawick and Beavin, 1967), paraphrasing (Grondin et al., 2019), interpreting (Sharma et al., 2020), agreement and acknowledgment, exploration through asking of genuine questions (Amjad et al., 2023) and projection (Bylund and Makoul, 2002).

Cognitive empathy failure. The category *cognitive empathy failure* is defined as the observer not actively putting themselves in the target's position and instead offering their subjective interpretation of the situation as a fact. Behavioral indicators include expressions of disbelief, questioning, doubt, denial, and disagreement (Bylund and Makoul, 2002; Pounds et al., 2018), offering a personal opinion as a fact, and blaming.

Affective empathy success. The category *affective empathy success* is focused on the observer's emotional experience of the target's situation. The category is defined as the observer expressing the experience of an emotional state congruent or similar to that of the target. The observer reacts to the target compassionately with emotional warmth and concern. Behavioral indicators are the expression of matching emotions (Clark et al., 2018), validation (Bylund and Makoul, 2002), appreciation (Amjad et al., 2023), praise and apologies, and offers for help and support(Pounds, 2011; Sharma et al., 2020).

Affective empathy failure. *Affective empathy failure* describes instances where the observer mentions their emotional state that is not congruent to that of the target's state. They react to the target with emotional coldness and harshly with no concern, discounting the target's situation. Behavioral indicators of empathy failure include orders, commands, unsolicited advice, dismissal, invalidation, coldness, and the communication of incongruent emotions (Amjad et al., 2023; Pounds et al., 2018).

Absence of empathy. The label *absence of Empathy* was introduced to be able to differentiate expressions that are neutral from expressions of empathy failure. It is used in cases when the observer responds merely factually rather than emotionally. Expressions do not contain elements that constitute empathy failure or empathy success.

Annotation process. We establish universal coding rules for the use of the annotating scheme. The coding unit is allowed to range from one word up to a compound of two sentences, depending on the semantic meaning. Moreover, only one label can be assigned to a coding unit. The coding procedure followed O'Connor and Joffe (2020) recommendations to ensure reliability. Two native German speakers coded our dataset using our annotation scheme. One of them was a master's student of psychology with a bachelor's degree in psychology, and the other was a master's student of organizational psychology who held a degree in business administration. Thus, the annotators were experts in relevant domains. As a first step, the first annotator applied the coding scheme to the first 25 percent of the LeadEmpathy dataset (ID1 - ID130). In the process, anchor examples for each of the coding system's indicators were added to the coding scheme. Next, the second annotator also coded the first 25 percent of the material (ID1 - ID130). After both had finished their annotations, inter-annotator agreement for each of the categories was calculated. This approach was based on the assumption that the randomly chosen portions (e.g. first 25 percent) of the data accurately represent the dataset (O'Connor and Joffe, 2020). Using Krippendorff's α (Krippendorff, 1980) as a metric, the agreement was acceptable for the success categories but initially insufficient for the failure categories. Both annotators together reassessed points of disagreement and the coding rules were refined. Annotators 1 and 2 proceeded to code the next 25 percent of the material (ID 131 - ID 260). Once completed, Krippendorff's α for the inter-annotator agreement was calculated again, revealing very good results (See section 4.2 for statistics). Upon achieving this reliability, the first annotator proceeded to code the remaining parts of the data independently.

4. The Dataset

4.1. Data Collection and Filtering

The data was collected in an experimental vignette study conducted in January 2023, as part of a research project that aimed to investigate the use and acceptance of AI among leaders. The participants were recruited via LoopsterPanel, a German participant recruitment tool with over 60,000 registered users. Since our study simulates the organizational context and therefore requires a basic understanding of work processes, we specifically recruited participants who had more than two years of work experience. The survey also included two attention checks, one being a question about the case and the other the instruction to choose a specific response option for an item, which resulted in immediate disqualification.

In the experiment, participants were presented with a scenario in which they were told to imagine themselves in a leadership function of a customer service department. In a series of unfortunate circumstances, one of the subordinates had made a severe mistake when processing an order, which is why an important customer was lost. The situation was purposefully portrayed as ambiguous in regard to who was at fault. In short, the employee should not have processed the customers' order without consulting the leader but did so anyway due to time urgency, after consulting other colleagues. However, the leader (in whose role the participants were supposed to put themselves) was not available throughout the whole day due to meetings. The scenario presented an empathetic opportunity, as the employee had good intentions and knew that there was time pressure to process the order.

After giving the participants some time to read and understand the circumstances, they were instructed to write an email to their subordinates (whom we named Mr. Thiele) and react to the situation at hand. The experimental manipulation consisted of feedback that followed the email draft, informing participants that either an artificial intelligence or a human had rated their email as not being sufficient in the amount of empathy conveyed in the email. Participants were then given the opportunity to either modify their initial email, write a new email or keep the original draft.

The experiment resulted in a total sample of N = 522 participants, with two emails per participant (email 1, email 2). In the filtering process, 17 participants were excluded from the data right away for not having made an entry for at least one of the two emails. A further 120 of the participants' submissions were labeled as invalid for a) being in a language other than German or English, b) non-compliance with the scenario and instead directly addressing the leading researcher and commenting about the experiment, c) misunderstanding the vignette and writing the email from someone else's point of view or addressing it to the wrong person, or d) writing nonsense. This resulted in a final sample of N= 385 participants, of which 183 were in the AI feedback condition and 202 were in the human expert feedback condition. This resulted in a dataset of 770 valid emails with an average length of 60 words (email 1: 53 words, email 2: 67 words).

In terms of demographics, 2.6% of participants were within the age range of 18-24 years, 20.8% were within 25-34 years, 26.0% were within 35-44 years, 24.2% were within 45-54 years and 26.5 were over the age of 54. The gender distribution of the participants was as follows: 52% were male, 47% were female, and 0.3 % diverse (1 participant). The work experience of the participants varied, with most 71.4% having more than 10 years of experience, 17.1% having 5-10 years of experience, and 11.4% having 2-5 years of experience. 41.4% of the participants were in a leadership position, 54.0 % were employees without a leadership position and 4.2 % chose the option "other".

4.2. Dataset Statistics

Inter-annotator agreement. Inter-annotator agreement was assessed to measure the reliability of our annotated dataset, using Krippendorff's α (Krippendorff, 1980) as our primary metric. Krippendorff's α = .80 constitutes very good agreement, while α > .667 represents an acceptable result (Krippendorff, 1980). We calculated Krippendorff's α on the e-mail level for each of our main categories (cognitive empathy failure and success, affective empathy failure and success) and higher-order categories (empathy success, empathy failure, and total empathy score) for 25% of the data. We present the results in Table 3, together with the frequencies of all categories. The threshold of α = .80 was met by all categories except for affective empathy success, which is one percentage point below. The results overall indicate a high level of agreement among annotators, which demonstrates that our annotated dataset and annotation scheme can be reliably used for subsequent analysis.

Annotations. We calculated measures of central tendency and dispersion of the final empathy score to get an overview of the range of empathy levels represented in the emails within our dataset. The descriptive statistics are shown in Table 2. The results indicate that the minimum score in email 1 is two points lower than in email 2, while the maximum score in email 2 is two points higher. Moreover, on average, email 2 received a higher empathy score than email 1. Overall, the score ranges from -4 to 7.

	Empathy Score		
	Email 1	Email 2	All Emails
Min.	-4	-2	-4
Max.	5	7	7
Median	1	2	1
Mean±SD	1.5 ± 0.7	1.8 ± 1.8	1.2 ± 1.7

Table 2: Measures of central tendency and dispersion of the empathy scores

Table 3 displays the frequency with which the four categories were assigned in the material. It contains the frequencies of categories in email 1, email 2, and the total emails. In email 1, the category cognitive empathy success received the highest frequency of assignments (229 instances), closely followed by affective empathy success (220 instances). In email 2, affective empathy success was the most frequently assigned category (416 instances), with cognitive empathy success being the second most frequently assigned category (386 instances).

In terms of failures, in both emails, instances of cognitive empathy failure occurred more often than affective empathy failure. Notably, in email 2, both success categories were assigned more often, and both failure categories received fewer assignments, again indicating that email 2 was rated as more empathetic.

5. Case Study: Empathy Detection

We present a case study on empathy detection with baseline models and briefly discuss insights from these experiments. For the scope of this work, we perform three detection tasks related to the overall empathy score, leaving experiments with finer-grained tasks for future work:

Binary classification (Bin). We classify empathy scores into two categories: 'Low,' which includes scores between -3 and 1 (i.e., empathy failures and low empathy), and 'High,' which encompasses scores from 2 to 7 (i.e., successful empathy). These score ranges create a fair balance between the two classes (see Fig. 2).

Multiclass classification (Multi). We predict the empathy scores as discrete classes for all ten overall empathy scores observed in our data. Figure 3 shows the class distribution.

Multiclass regression (RegCLS). We predict the empathy scores by regression as opposed to discrete classes. For our analysis, we convert the predicted scores to the nearest discrete values,

	CE		AE		CE+AE		Empathy score
	failure	success	failure	success	failure	success	
Email 1	117	229	50	220	167	449	-
Email 2	92	386	34	416	126	805	-
Total	209	618	84	636	293	1254	
Krippendorff's α	0.84	0.85	0.80	0.79	0.85	0.88	0.91

Table 3: Category frequencies and inter-coder reliability (CE = cognitive empathy, AE = affective empathy)

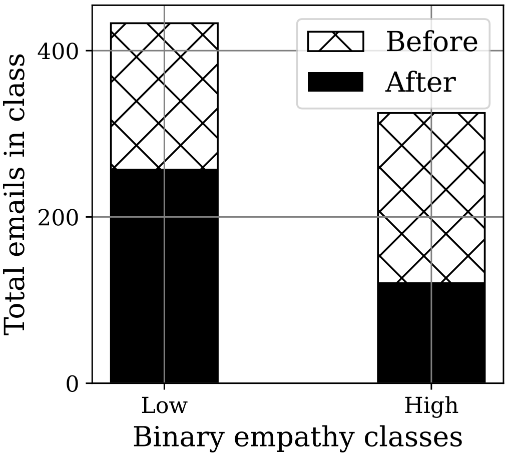


Figure 2: Distribution of binary empathy classes.

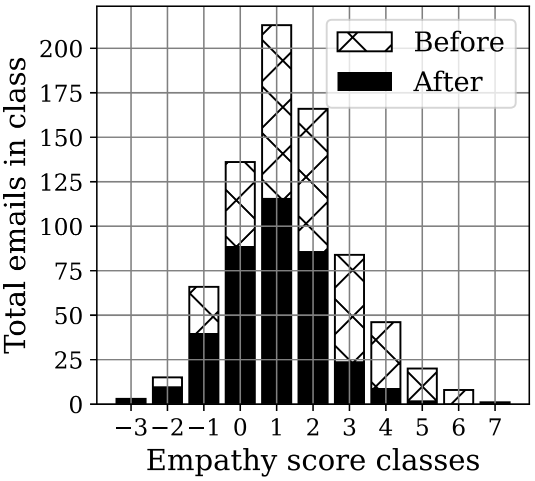


Figure 3: Distribution of overall empathy scores.

thus formulating this approach as multiclass classification.

For these experiments, we treat the before and after emails as separate instances. We evaluated an SVM[2,3] with n-gram and LIWC (Boyd et al., 2022) feature sets, and a BERT-based model by finetuning `bert-base-german-cased`.[4] As the scope of this work is primarily the development and publication of the LeadEmpathy dataset, we limited the complexity of these case study experiments by refraining from in-depth parameter and feature set exploration, reserving these investigations for future research.

Task	Best model	Majority class	Prec.	Rec.	F1
Bin	SVM	57.1	81.7	81.8	81.7
Multi	BERT	28.1	45.2	47.1	45.7
RegCLS	BERT	28.1	50.1	50.4	49.9

Table 4: Results for models with the best performance by weighted F1 score on each task.

We train and evaluate in a 10-fold cross-validation setup, stratifying the labels. All experiments were run on the same samples per fold for comparison. We present the performance results of the best performing models (SVM with unigrams for Bin and BERT for Multi and RegCLS) in Table 4. Each outperformed the majority class baselines significantly.

Though these are simple models, they demonstrate the capacity to detect a wide range of empathy scores, and distinguish between empathic failures and successes. In the case of RegCLS, we also evaluated the performance by regression metrics before converting the predicted scores to classes. The model predictions had a Pearson correlation with the actual scores of 0.816 and mean squared error of 0.883. Thus, on average, the incorrect predictions were only less than one point away from the actual score out of a range of ten points. Thus, the model is quite effective at predicting the overall empathy scores. In future work, we will expand the experiments to include diverse tasks for predicting cognitive and emotional empathy. Furthermore, we can integrate the particular text segments that the annotators marked pertaining to their particular empathy labels and evaluations.

[2] LinearSVC from https://scikit-learn.org/.

[3] We performed preliminary experiments with naive bayes, logistic regression, and other classifiers and observed SVM performed best in all tasks in our experimental setup.

[4] https://huggingface.co/bert-base-german-cased

6. Conclusions and Future Work

In this work, we introduce an innovative theory-based annotation scheme for discerning both affective and cognitive empathy as well as empathy successes and failures in written leadership communication. The coding scheme, developed by experts in business psychology, provides a valuable framework for analyzing empathetic aspects of leadership texts. Moreover, we introduce the LeadEmpathy dataset, offering expert-annotated empathy language resource in German, in the novel domain of written leadership communication. Upon independent coding, two annotators reached a substantial inter-annotator agreement for all categories, indicating the reliability of the approach and the quality of the annotations. The dataset is designed to facilitate the modeling of empathy in written leadership communication, addressing a critical gap in the availability of such datasets and contributing to the broader understanding of empathy in this domain. This work therefore makes contributions toward both the need for comprehensive empathy resources and informed frameworks and measurement approaches that are reliable in the face of the task's subjectivity. We encourage the use of our annotation scheme and dataset for further explorations of empathy. Specifically, we suggest that our work can be a valuable resource for scholars aiming to understand the behavioral dimension of empathy and compare aspects of empathetic communication across domains and languages.

Developing robust NLP models. Based on our work, computational analyses can be used to further explore and theoretically refine the behavioral dimensions of empathy. For example, a better understanding of the nature of empathy can be gained through leveraging NLP models to identify latent traits that distinguish empathy successes and failures. Computational analyses provide a data-driven and context-sensitive approach to understanding as well as quantifying how empathy is expressed. While we demonstrated a baseline model performing simple tasks for this dataset, there are many other NLP modeling tasks to investigate. Future experiments can leverage the labeled span segments for improvements to detection and recognition models. Of particular interest is the development of models capable of distinguishing elements of affective and cognitive empathy. Multilabel classification models can be developed for multitask predictions of AE and CE. Furthermore, the multidimensionality of the empathy construct employed can support the development of more robust NLP models empathic language and exchanges. Future investigations to support this objective could involve studying the transferability of empathic signals from this domain to others, and vice versa.

Cross-domain and cross-language comparisons. In further studies, our dataset can be leveraged to compare verbal expressions of empathy across domains and languages. For example, therapy conversations may involve more affective empathy compared to leadership interactions due to varying social norms and the nature of the conversations. Moreover, the tasks, responsibilities, and role of a leader differ significantly from those of a therapist. While therapists primarily seek to improve the well-being of the patient, the leader must primarily act in the interest of the organization while keeping the well-being of the employees in mind. It might also be that business etiquette differs in organizational settings across different cultures and languages, e.g. in some countries it might be the norm to have stronger organizational hierarchies which might not allow expressions of empathy as frequently as in cultures with flatter hierarchies as the norm. Altogether, more comparative studies are needed to develop a nuanced understanding of empathetic expression, representing a promising future research area.

7. Acknowledgements

We sincerely thank our research assistant Katharina Neufeld for her high quality work on the annotations. This work has been supported by the German Federal Ministry of Education and Research (BMBF) as a part of the Junior AI Scientists program under the reference 01-S20060 and the state of North Rhine-Westphalia as part of the Lamarr Institute for Machine Learning and Artificial Intelligence. Moreover, this work is part of the project on responsible algorithmic decision-making at the Centre for Responsible Digitality (ZEVEDI).

8. Ethics statement

This research was reviewed and approved by an institutional Ethics Committee for human sciences of the Faculty of Human Sciences at the University of Kassel, Germany.

9. Bibliographical References

Bushra Amjad, Muhammad Zeeshan, and Mirza Omer Beg. 2023. Emp-eval: A framework for measuring empathy in open domain dialogues.

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

Bradford S. Bell, Kristie L. McAlpine, and N. Sharon Hill. 2023. Leading virtually. *Annual Review of Organizational Psychology and Organizational Behavior*, 10(1):339–362.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

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.

Carma Bylund and Gregory Makoul. 2005. Examining empathy in medical encounters: An observational study using the empathic communication coding system. *Health communication*, 18:123–40.

Carma L Bylund and Gregory Makoul. 2002. Empathic communication and gender in the physician–patient encounter. *Patient Education and Counseling*, 48(3):207–216.

Malissa Clark, Melissa Robertson, and Stephen Young. 2018. "i feel your pain": A critical review of organizational research on empathy. *Journal of Organizational Behavior*, 40.

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

Coral J Dando, Gavin E. Oxburgh, and Gavin E. Oxburgh. 2016. Empathy in the field: Towards a taxonomy of empathic communication in information gathering interviews with suspected sex offenders. *European Journal of Psychology Applied to Legal Context*, 8:27–33.

Mark Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personalilty and social psychology*, 44:113–126.

Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Incorporating politeness across languages in customer care responses: Towards building a multi-lingual empathetic dialogue agent. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4172–4182, Marseille, France. European Language Resources Association.

Frédéric Grondin, Anna Lomanowska, and Philip Jackson. 2019. Empathy in computer-mediated interactions: A conceptual framework for research and clinical practice. *Clinical Psychology Science and Practice*, page e12298.

Mahshid Hosseini and Cornelia Caragea. 2021. Distilling knowledge for empathy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3713–3724, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ned Kock, Milton Mayfield, Jacqueline Mayfield, Shaun Sexton, and Lina M. De La Garza. 2019. Empathetic leadership: How leader emotional support and understanding influences follower performance. *Journal of Leadership & Organizational Studies*, 26(2):217–236.

Klaus Krippendorff. 1980. *Content analysis: An introduction to its methodology*. SAGE Publications.

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.

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.

Rubina Mahsud, Gary A. Yukl, and Gregory E. Prussia. 2010. Leader empathy, ethical leadership, and relations-oriented behaviors as antecedents of leader-member exchange quality. *Journal of Managerial Psychology*, 25:561–577.

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.

Shannon L. Marlow, Christina N. Lacerenza, and Eduardo Salas. 2017. Communication in virtual teams: a conceptual framework and research agenda. *Human Resource Management Review*, 27(4):575–589. Virtual Teams in Organizations.

Philipp Mayring. 2014. *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. Klagenfurt.

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.

Cliodhna O'Connor and Helene Joffe. 2020. Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, 19:1609406919899220.

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.

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

Gabrina Pounds, Daniel Hunt, and Nelya Koteyko. 2018. Expression of empathy in a facebook-based diabetes support group. *Discourse, Context Media*, 25:34–43.

Niels Van Quaquebeke and Fabiola H. Gerpott. 2023. The now, new, and next of digital leadership: How artificial intelligence (ai) will take over and change leadership as we know it. *Journal of Leadership & Organizational Studies*, 30(3):265–275.

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.

Simone G. Shamay-Tsoory. 2011. The neural bases for empathy. *The Neuroscientist*, 17(1):18–24. PMID: 21071616.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, WWW '21, page 194–205, New York, NY, USA. Association for Computing Machinery.

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.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20, 1st virtual meeting. Association for Computational Linguistics.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, 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.

Christian Tuschner, Jeanine Kirchner-Krath, Jan Bings, Marvin Schwenkmezger, Manuel Etzkorn, and Harald von Kortzfleisch. 2022. Leading in the digital age: A systematic review on leader traits in the context of e-leadership.

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.

Thiemo Wambsganss, Matthias Soellner, Kenneth R Koedinger, and Jan Marco Leimeister. 2022. Adaptive empathy learning support in peer review scenarios. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Paul Watzlawick and Janet Beavin. 1967. Some formal aspects of communication. *American Behavioral Scientist*, 10(8):4–8.

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, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2021. Towards low-resource real-time assessment of empathy in counselling. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 204–216, Online. Association for Computational Linguistics.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 133–147, Online. Association for Computational Linguistics.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

10. Appendix

Figure 4 shows examples of the annotation scheme applied to the original German emails. Table 5 shows the full annotation scheme.

Example containing success and failure (ID 366, Email 1)

Sehr geehrter Herr Thiele,

ich bin mir der Schwierigkeit der Entscheidung bewusst, die Sie treffen mussten und **schätze es sehr, dass Sie im Team nachgefragt haben.** **Dennoch entschuldigt dies in keinster Weise den Fehler, den Sie bei der Auswahl der Geräte gemacht haben.** Über weitere Entscheidungen diesbezüglich werde ich sie zeitnah informieren und verbleibe bis dahin (...). Mit freundlichen Grüßen.

Example that only contains failure (ID 180, Email 1)

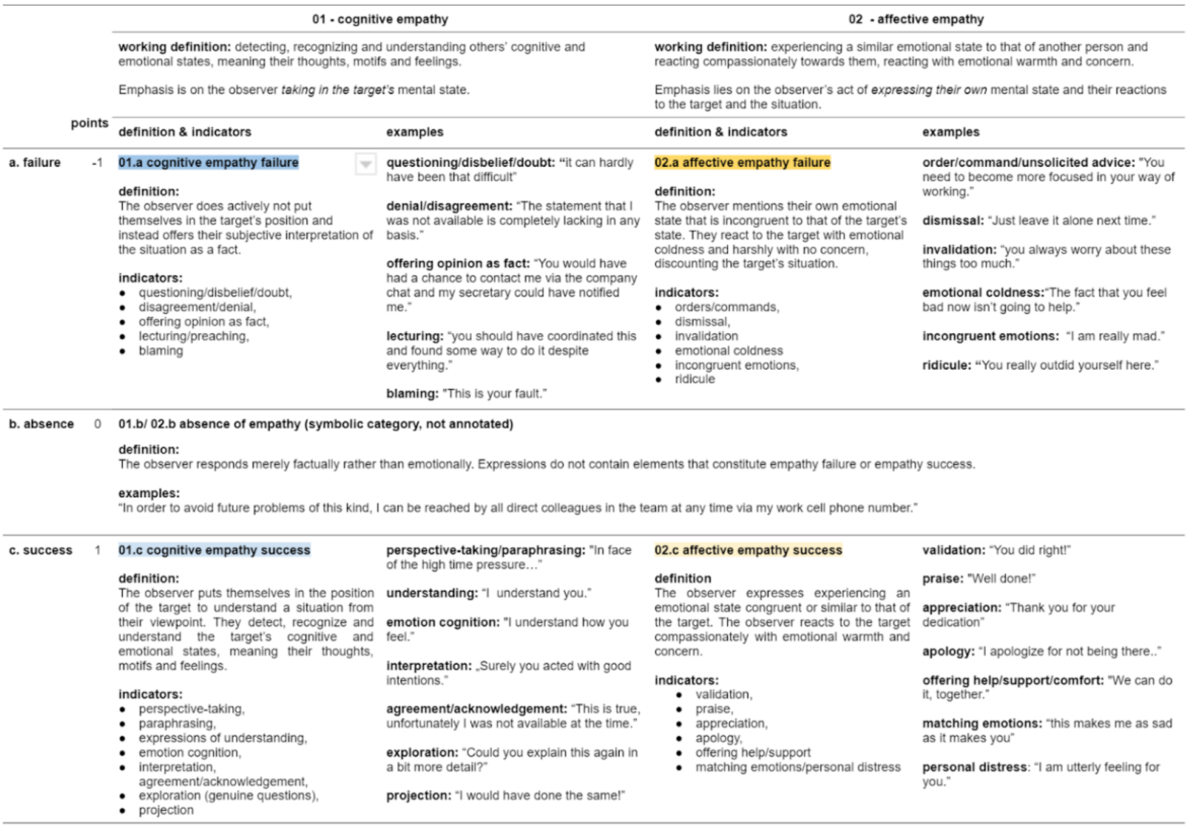
Ich war mindestens per Email zu erreichen und **erwarte, dass Sie die Verantwortung übernehmen**.

Example that only contains success (ID 487, Email 2)

Hallo Herr Thiele, **vielen Dank für Ihre prompte und ausführliche Antwort**. **Ich versichere Ihnen, dass Sie sich wegen Ihres Fehlers nicht schlecht fühlen müssen**, denn **wer arbeitet macht Fehler.** **Ich stehe in dieser Sache hinter Ihnen** und denke es ist an der Zeit unseren Kunden klar zu machen, dass **Fehler passieren. Diese Fehler tun uns sehr leid, aber es nur allzu menschlich ist.**. Selbstverständlich werden wir den finanziellen Schaden unseres Kunden ersetzen. **Sie haben getan, was in ihren und meinen Augen notwendig war**, und **dafür danke ich Ihnen sehr.**

Key cognitive empathy: **success** **failure** affective empathy: **success** **failure**

Figure 4: Examples from the dataset demonstrating the application of the coding scheme in the original German.



		01 - cognitive empathy		02 - affective empathy	
		working definition: detecting, recognizing and understanding others' cognitive and emotional states, meaning their thoughts, motifs and feelings. Emphasis is on the observer *taking in the target's* mental state.		**working definition:** experiencing a similar emotional state to that of another person and reacting compassionately towards them, reacting with emotional warmth and concern. Emphasis lies on the observer's act of *expressing their own* mental state and their reactions to the target and the situation.	
	points	**definition & indicators**	**examples**	**definition & indicators**	**examples**
a. failure	-1	**01.a cognitive empathy failure** **definition:** The observer does actively not put themselves in the target's position and instead offers their subjective interpretation of the situation as a fact. **indicators:** • questioning/disbelief/doubt, • disagreement/denial, • offering opinion as fact, • lecturing/preaching, • blaming	**questioning/disbelief/doubt:** "it can hardly have been that difficult" **denial/disagreement:** "The statement that I was not available is completely lacking in any basis." **offering opinion as fact:** "You would have had a chance to contact me via the company chat and my secretary could have notified me." **lecturing:** "you should have coordinated this and found some way to do it despite everything." **blaming:** "This is your fault."	**02.a affective empathy failure** **definition:** The observer mentions their own emotional state that is incongruent to that of the target's state. They react to the target with emotional coldness and harshly with no concern, discounting the target's situation. **indicators:** • orders/commands, • dismissal, • invalidation • emotional coldness • incongruent emotions, • ridicule	**order/command/unsolicited advice:** "You need to become more focused in your way of working." **dismissal:** "Just leave it alone next time." **invalidation:** "you always worry about these things too much." **emotional coldness:**"The fact that you feel bad now isn't going to help." **incongruent emotions:** "I am really mad." **ridicule:** "You really outdid yourself here."
b. absence	0	**01.b/ 02.b absence of empathy (symbolic category, not annotated)** **definition:** The observer responds merely factually rather than emotionally. Expressions do not contain elements that constitute empathy failure or empathy success. **examples:** "In order to avoid future problems of this kind, I can be reached by all direct colleagues in the team at any time via my work cell phone number."			
c. success	1	**01.c cognitive empathy success** **definition:** The observer puts themselves in the position of the target to understand a situation from their viewpoint. They detect, recognize and understand the target's cognitive and emotional states, meaning their thoughts, motifs and feelings. **indicators:** • perspective-taking, • paraphrasing, • expressions of understanding, • emotion cognition, • interpretation, agreement/acknowledgement, • exploration (genuine questions), • projection	**perspective-taking/paraphrasing:** "In face of the high time pressure…" **understanding:** "I understand you." **emotion cognition:** "I understand how you feel." **interpretation:** „Surely you acted with good intentions." **agreement/acknowledgement:** "This is true, unfortunately I was not available at the time." **exploration:** "Could you explain this again in a bit more detail?" **projection:** "I would have done the same!"	**02.c affective empathy success** **definition** The observer expresses experiencing an emotional state congruent or similar to that of the target. The observer reacts to the target compassionately with emotional warmth and concern. **indicators:** • validation, • praise, • appreciation, • apology, • offering help/support • matching emotions/personal distress	**validation:** "You did right!" **praise:** "Well done!" **appreciation:** "Thank you for your dedication" **apology:** "I apologize for not being there.." **offering help/support/comfort:** "We can do it, together." **matching emotions:** "this makes me as sad as it makes you" **personal distress**: "I am utterly feeling for you."

Table 5: The full annotation scheme.