Modeling Empathetic Alignment in Conversation

Jiamin Yang
University of Chicago
jiaminy@uchicago.edu

David Jurgens
University of Michigan
jurgens@umich.edu

Abstract

Empathy requires perspective-taking: empathetic responses require a person to reason about what another has experienced and communicate that understanding in language. However, most NLP approaches to empathy do not explicitly model this alignment process. Here, we introduce a new approach to recognizing alignment in empathetic speech, grounded in Appraisal Theory. We introduce a new dataset of over 9.2K span-level annotations of different types of appraisals of a person's experience and over 3K empathetic alignments between a speaker's and observer's speech. Through computational experiments, we show that these appraisals and alignments can be accurately recognized. In experiments in over 9.2M Reddit conversations, we find that appraisals capture meaningful groupings of behavior but that most responses have minimal alignment. However, we find that mental health professionals engage with substantially more empathetic alignment.

1 Introduction

Empathy is a key aspect of successful clinical health conversations (Hojat et al., 2013; Raab, 2014). In general, empathy involves an emotional component, where a listener resonates with the emotional tone of a speaker, and a cognitive component, conveying the listener understands the speaker (Hatfield et al., 2011). Underlying both of these components is the perspective-taking by the listener to mirror the experience of the speaker, or as Mahrer (1997) describes it, "being aligned is another way of being empathic." While past computational work on empathy has measured how empathetic messages can be, we still understand little about what aligns the language and perspective. Here, we examine empathy as an alignment task, studying therapeutic conversations on Reddit.

Given the importance of empathy, particularly in the clinical setting, NLP methods have attempted to model the relative level of empathy in replies (Sharma et al., 2020; Omitaomu et al., 2022). Better models for recognizing empathy are aimed to help support generating more empathetic responses (e.g., Sharma et al., 2021; Welivita et al., 2023). However, as Lahnala et al. (2022) note, many of these works focus only on the emotional mirroring component of empathy, rather than its cognitive component of perspective taking, and none explicitly model the alignment between the speaker, known as the *Target*, and listener, as the *Observer*.

Here, we introduce a new dataset and computational models for studying empathetic alignment in conversation. To quantify alignment, our work draws on the Appraisal Theory (Wondra and Ellsworth, 2015), which describes six aspects of how a person may experience a situation, e.g., describing its pleasantness or how much control they had, and encompasses both cognitive and emotional components. This scheme gives us a fine-grain labeling of both what is described and how the person feels. Because both the Target and Observer can appraise the same content differently, this view provides critical insight for understanding whether the two are aligned.

This paper offers the following four contributions to the study of empathy in NLP. First, we introduce ALOE, a new dataset, of therapeutic Reddit conversations labeled with 9,284 appraisals from both the Target and Observer and 3,262 alignments between the Target and Observer. Our dataset goes beyond theory to introduce new categories that model common types of aligned spans. Second, in experiments, we show that appraisals can be accurately recognized and that the alignment between appraisals can be recognized, though we show that both are challenging tasks. Third, in analyses on the appraisals and alignments of 2.3M posts and 8.9M comments, we show that appraisals meaningfully capture differences in how individuals experience distressing situations and in how others reply—but that the dominant form of align-

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3127–3148

ment is to reply with advice, rather than a matched appraisal. Fourth, in comparisons between mental health professionals and laypeople on Reddit, professionals have much higher alignment with Targets; but, as seen in clinical settings, both professionals and laypeople decrease in their levels of alignment as they become more experienced.

2 Empathy in Therapeutic Settings

Empathy has been an important concept in social, personality, and clinical psychology (Davis, 2018; Eisenberg et al., 2013; Batson et al., 1981; Hall et al., 2021b). Though being diversely defined, the most discussed aspects are emotional empathy and cognitive empathy (Cuff et al., 2016). Emotional empathy focuses on the vicarious sharing of emotion, while cognitive empathy relates to mental perspective-taking (Smith, 2006; Shamay-Tsoory, 2011; Blair, 2005). In other words, emotional empathy is expressed as "I feel what you feel", and cognitive empathy is more commonly recognized as "I understand what you feel" (Healey and Grossman, 2018).

Empathetic conversation is thought to play an important role in the development of social relationships (Hoffman, 2001), and mental health professionals are taught to develop empathetic skills (Toombs, 2001; Moudatsou et al., 2020), to improve patient outcomes and experiences. Central to these empathetic conversations is the alignment between what a Target is feeling and confirmation that the Observer's mental model of the Target matches these feelings; explicit expressions of this alignment are important for a Target to experience an Observer's response as empathetic (e.g., Thwaites and Bennett-Levy, 2007; Vyskocilova et al., 2011; Watson, 2016). While related to concepts like "active listening" or "reflective listening," this type of speech requires a communication of the Observer's theory of mind to show that they have understood what the Target has experienced, rather than just repeating parts of what a Target has said.

Individuals seeking mental health support increasingly turn to social media (Hanley et al., 2019). Compared with traditional therapy sessions, the observers are no longer guaranteed to be trained professionals and the interactions are largely text-only. Given abundant data and unique features, empathy in online communities becomes a valuable subject for active research (Naslund et al., 2016), including comparisons of defining and expressing empathy between laypeople and professionals (Hall et al., 2021a; Lahnala et al., 2021).

Within NLP, significant work has been done in predicting empathy (Guda et al., 2021; Vasava et al., 2022), analyzing empathetic expressions and behaviors (Sharma et al., 2020; Zhou and Jurgens, 2020), and facilitating empathetic conversations (Sharma et al., 2021; Xie and Pu, 2021; Zeng et al., 2021; Zhu et al., 2022). However, issues have been pointed out where empathy definitions are absent or abstract, and emotional empathy is overemphasized, while cognitive empathy is often absent or minimized (Lahnala et al., 2022).

NLP models for recognizing empathy typically treat empathy as a classification or regression task. However, this introduces a gap: in clinical settings, speaking with empathy is often viewed as *aligning* the Observer's speech to the Target's, yet we lack methods for how to explicitly identify this alignment. Our work directly addresses this gap by recognizing cognitive and emotive appraisals (Smith et al., 2010; Lamm et al., 2007; Wondra and Ellsworth, 2015) and measuring empathy in terms of the degree of Observer alignment with a Target's situation and appraises it in the same way.

3 A Dataset of Empathetic Appraisals

To facilitate research on cognitive and emotional empathy, we introduce a new dataset of Target and Observer pairs, ALOE (**Al**igment **o**f **E**mpathy), annotated for how each appraised the Target's situation and which appraised passages are aligned.

3.1 Data Source

Data was drawn from Reddit, which hosts a diverse range of communities focused on mental, emotional, and social support (De Choudhury and De, 2014; Gkotsis et al., 2016). Support typically occurs in two settings. Most commonly, an individual in need of support with make a post describing their situation, and then others may reply in comments to the post; additionally, a user may comment in a conversation thread that solicits a supportive discussion, e.g., a weekly post requesting such comments. Candidate data for annotation was selected from all post-comment pairs and comment-reply to those posts in 35 English-language subreddits (Appendix A) from 2019-01 to 2021-06. This collected 28,018 post-comment and 1367 comment-comment candidate pairs for annotation.

Not all content in these communities relates to

empathy, e.g., off-topic conversations or posts from moderators. To focus specifically on empathy-related content, we pre-filter data using the models of Zhou and Jurgens (2020); their models identify content relating to distress, whether a reply is condolence, and an ordinal measure of the empathy of a reply. Details of these classifiers are in Appendix B. We retain only annotation candidates where (1) the post was classified as distress and the reply as condolence and (2) the empathy rating for the reply was $\geq$2, on a scale from $[1, 5]$. This latter constraint was designed to prioritize content likely to have empathetic appraisals, as the majority of replies are low-empathy. Finally, we discard pairs where the Target contained ≥ 3 uses of "you" to avoid cases where the Target was itself a response to other distress posts or comments. In total, 29,385 Target-Observer pairs were collected.

3.2 Annotation Task and Process

Our annotation process consisted of extensive pilot work to develop annotation guidelines and multiple rounds of annotation and discussion.

Tasks Two annotation tasks were performed. The first asked annotators to highlight spans of the Target's and Observer's texts that matched one of 9 categories. Here, we include the six appraisal categories proposed by Wondra and Ellsworth (2015), described in Appendix C.1. Our initial pilot work identified three other categories that warranted annotation. Target often includes some description of the situation that is neutral with respect to their appraisal, which we label as *Objective Experience*or they may actively ask for advice from others (*Advice*). Observers, in turn, may also share similar experiences (*Objective Experience*), provide suggestions or advice (*Advice*), or use sympathetic tropes such as "I'm sorry for your loss" (*Trope*). We include these additional span types as (1) they each reflect a common category of response type seen in everyday language—not just Reddit, (2) their inclusion helps annotators distinguish each construct from the appraisals, and (3) they offer a new way to model empathetic alignment beyond appraisals and provide more structure for understanding the lived experiences of how people receive social support, e.g., by identifying how others empathize (or struggle to) in their responses. Examples spans of these appraisals are shown in Appendix Table 6.

Annotators were allowed to highlight spans of varied length, from clauses to multiple sentences, depending on how the individual wrote. Annotators were instructed to label a passage with only a single span type; if a sentence contained multiple span types, each should be marked separately.

The second task had annotators align the spans between Target and Observer. Annotators were shown all labeled spans of the first phase and asked to identify any pairs where the Observer's span references a Target. An Observer span was allowed to be aligned to multiple Target spans, as often the Observer attempts to summarize and synthesize what the Target has said in their response. Full annotation instructions for both tasks are described in Appendix C.1

Annotation Process The annotation process is divided into two phases: annotating the spans of appraisals, and annotating the alignment of spans between Target and Observer. Due to the complexity of the task, annotators were recruited in person to receive training. Five annotators participated and went through six hours of training using the annotation codebook reported in Appendix C.1. Following training, annotators worked and met weekly to discuss controversial annotations across annotators. Annotators used a custom web interface to annotate (Appendix C.3), which also allowed them to take notes on specific instances they wanted to discuss, which were used to improve the codebook when applicable. Phase 1 annotations were completed in batches of 634 instances.

Phase 2 alignment annotations were completed by 4 annotators who were also involved in producing the labels for Phase 1. Annotators used a custom web interface shown in Appendix Figure 9 following a separate codebook for deciding when spans were aligned.

Adjudication Process In both phases, following each batch's completion, annotators participated in a review and adjudication process where all were allowed to compare their annotations with others, leave comments on why they labeled certain appraisals, and make changes to annotations of their own will. This process was designed to let annotators have access to different mindsets from others, as interpreting appraisals can be subjective based on one's own way of understanding the situation. Once Phase 1 annotation was complete, all remaining disagreements were resolved by one expert annotator prior to starting Phase 2. Following the completion of Phase 2, one expert annotator resolved all remaining disagreements on alignment.

Span Type	Target	Observer	Has alignment in Observer
Pleasantness	1059	487	522
Situational Control	744	278	268
Anticipated Effort	738	357	273
Self-other Agency	906	507	465
Certainty	798	541	393
Attentional Activity	223	40	74
Objective Experience	885	362	168
Advice	137	857	103
Trope	0	363	0

Table 1: Statistics of ALOE dataset.

Because of adjudication, we do not report IAA, as this is not a meaningful estimate of reliability. Annotating appraisals is challenging due to the perspective-taking required, and adjudication was essential for mutual conceptualizing and agreeing upon the likely appraisals in many cases. We describe the challenges later in Section 3.3.

Annotated Dataset Summary Annotators ultimately identified 9,284 spans across 636 Target-Observer pairs, with 3,262 alignments across spans. Table 1 shows the appraisal counts for both Target and Observer, and how many times a Target's appraisal was aligned with an Observer span.

3.3 Challenges in identifying appraisals

Three common themes in difficulties were encountered during annotation, described next.

Implicit Expressions Some emotions are inferential and implicit in the text. For example, a user may say "My cat died yesterday", which would be considered *Pleasantness* if we infer the likely emotion experienced. However, due to the distressing content, many such passages would be rated for inferred Pleasantness and so we opt to only rate explicit mentions of emotion.

Ambiguity The language of some spans was sufficiently ambiguous to elicit multiple appraisal types, e.g."Depression in relationships can be tough." The phrase "tough" could be interpreted with respect to emotion (*Pleasantness*) or the amount of effort needed *Anticipated Efforts*.

Descriptions of Attention Among all appraisal types, *Attentional Activity* was most difficult to distinguish due to the infrequency with which Targets explicitly focus on their surprise or focus of attention; instead, such language is used to indicate other types of appraisals that are more dominant in their salience, leading to its rarity in our data.

4 Classifying Appraisals and Alignment

Models were trained to identify spans of appraisals and to align spans between Target and Observer.

4.1 Appraisal Prediction

We first performed the task of automatically annotating appraisals in both Target and Observer. Due to its rarity, we excluded the *Attentional Activity* type from our model and set them to be *No Label* in this task. Most of the annotated spans were whole sentences, except the case where subsentences showed observable different appraisals, so we predicted at the sentence level, i.e. given a Target or Observer text containing l sentences: $\langle s_1, s_2, \ldots, s_l \rangle$, each $s_i, 1 \leq i \leq l$ is passed to the model independently to be predicted. When multiple appraisals were present, we selected the longer span in terms of characters and, when equal in length, arbitrarily broke ties. We combined data from Target and Observer when training models.

Classification models were trained starting from pre-trained language models (PLMs): BERT-large-uncased (Devlin et al., 2019), RoBERTa-large (Liu et al., 2019), SpanBERT-large-cased (Joshi et al., 2020), DeBERTa-v3-cased (He et al., 2023), sentence-transformers/all-MiniLM-L6-v2 (Wang et al., 2020). We also tested using a prompt-based models: OpenPrompt+BERT-large-uncased (Ding et al., 2021; Devlin et al., 2019), OpenPrompt+RoBERTa-large (Ding et al., 2021; Liu et al., 2019), OpenPrompt+T5-large (Ding et al., 2021; Raffel et al., 2020). Additional training details are reported in Appendix D.4. The baseline was set as the random prediction.

Results In general, prompt-based models performed better than PLMs, as shown in Table 2, and all models outperformed the baseline. Examining appraisal-level performance (Appendix Table 8), we saw that *Advice*, *Trope*, and *Objective Experience* were the easiest to classify, while *Anticipated Effort* as the lowest. However, classification performance was similar for most appraisal types, indicating the model was sufficiently effective to label data for large-scale analysis.

4.2 Alignment Prediction

Alignment prediction between Target and Observer was done using a Siamese Network (Bromley et al., 1993). The task is structured as, given a span of text from the Target and a span of text from the Observer, predict whether the Observer's appraisal is

	F1	Recall	Precision
random	0.11	0.11	0.11
majority	0.03	0.02	0.11
BERT	0.38	0.38	0.41
RoBERTa	**0.56**	0.56	0.57
SpanBERT	0.52	0.52	0.54
DeBERTa	0.55	0.55	0.56
MiniLM	0.49	0.50	0.51
OpenPrompt+BERT	0.53	0.53	0.55
OpenPrompt+ RoBERTa	**0.56**	**0.57**	0.58
OpenPrompt+ T5-large	**0.56**	0.56	**0.59**

Table 2: Appraisal model performance.

aligned with the Target's appraisal. Formally, given a pair of Target and Observer (T, O) with annotated appraisals/spans where $T = \{\text{span}_{t_1}, ..., \text{span}_{t_k}\}$, $O = \{\text{span}_{o_1}, ..., \text{span}_{o_l}\}$, the input data is $T \times O$ with label $Y \in \{0, 1\}^{k \times l}$. Because most pairs did not align, and the alignment between three pairs (*Advice*and *Objective Experience*, *Advice*and Pleasantness, Anticipated Effort and *Objective Experience*) does not exist or is extremely rare (fewer than 7 occurrences), when constructing the dataset, we omit those pairs and downsampled to a positive-negative ratio of 1:11. We tested using the all-MiniLM-L6-v2 or all-mpnet-base-v2 parameters to initialize the Siamese Network. The baseline was set as picking a random label with the empirical distribution of the dataset. We evaluated the performance using Binary F1 for is-aligned.

In addition to these two trained models, we also include two other baselines that focus just on text similarity: threshold classifiers trained on either the Jaccard Index of the words in the two passages or on a Siamese network with all-mpnet-base-v2 parameters. Both baselines allow us to test whether empathetic alignment is simply textual similarity, or, as theory predicts, a deeper alignment that goes beyond content.

Results Both Siamese network models were able to meaningfully identify alignment, as shown in Table 3, with the mpnet-base-v2 model performing best. Notably, both threshold-base baselines show that empathetic alignment requires more than text overlap or semantic similarity between two spans; while both baselines do attain high precision, they fail to recognize the majority of the cases where the Observer is aligning. However, alignment classification is a challenging task, with our best model only attaining a binary F1 of 0.46. In particular, the task requires significant social reasoning capabilities to understand how an Observer's speech is reflective of the Target's description, which provides significant room for improvement. Appendix table 12 shows examples highlighting the variety and subtly in determining alignment.

LM	recall	precision	F1
random	0.09	0.08	0.08
baseline: word overlap	0.01	0.55	0.02
baseline: all-mpnet-base-v2	0.02	0.62	0.04
fine-tuned all-MiniLM-L6-v2	0.41	0.42	0.41
fine-tuned all-mpnet-base-v2	**0.45**	**0.46**	**0.46**

Table 3: Alignment model performance.

4.3 Appraisal and Alignment Dataset

We applied our best model (OpenPrompt+RoBERTa) to predict appraisals in comments and posts in 91 subreddits relating to mental health and support (listed in Appendix A.2). For each post, we classified whether the post was about distress using the approach described in Section 3.1 and then labeled the appraisals for the post and all comments made under that post. After combining the consecutive sentences that were predicted to have the same appraisal, we passed them to all-mpnet-base-v2 for alignment prediction. We applied this pipeline of models to 2.3M posts and 8.9M comments, identifying 21.7M appraisals in Targets' posts or comments and 326.9M appraisals in Observers' comments. We used this dataset for all analyses.

5 Appraisal Behavior

Different types of distressing events may be more likely to evoke specific appraisals, such as (un)pleasantness for the loss of a loved one, or the effort involved to handle mental illness. The 91 communities in our data cover a range of possible situations and Stellar and Duong (2023) note that empathy must be understood in context, with responses that adapt to the circumstances. Here, we test whether individuals in these communities show regularity in how they appraise as Targets and, do Observers, in turn, vary the appraisals with which they respond.

Setup PCA is then run on a matrix of subreddits and their normalized distribution of appraisals across all their posts.

Results Communities were thematically clustered solely based on the relative distribution of appraisals (not content), shown in Figure 1 and Figure 2. For example, for Targets, clusters are seen

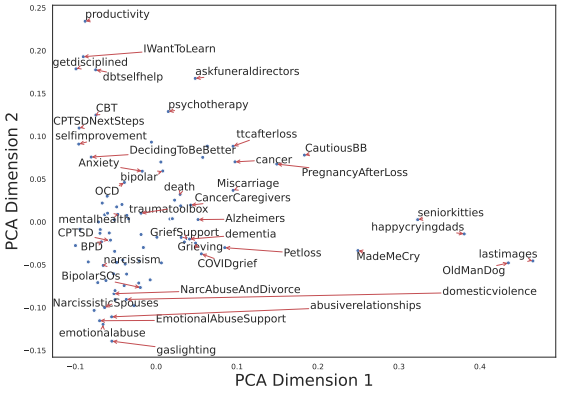

Figure 1: Subreddits arranged according to their distribution of Target appraisals.

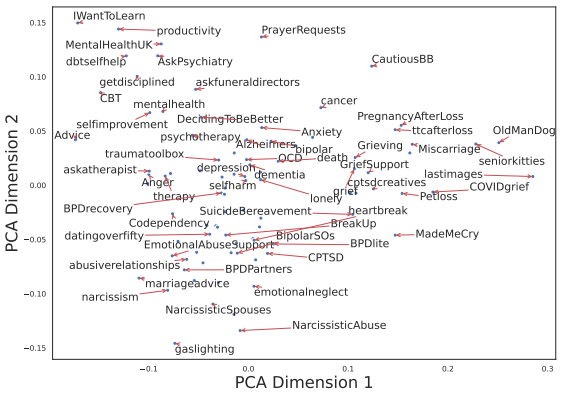

Figure 2: Subreddits arranged according to their distribution of Observer appraisals.

for communities focused on structured therapy and self-help modalities (top left), recipients of abuse (bottom left), and various topics of grief (center to center right). Similar clusters are also seen based on how Observers appraise in these subreddits. Although these subreddits all contain distressing situations, there is no a priori reason to expect that they should differ in how people categorize their lived experiences—many distressing situations could easily be described with any of the appraisals, yet individuals show regularity in how they rationalize and describe their thematically-similar experiences. This emergent grouping suggests that the related situations that targets find themselves in between these communities, while different, lead to similar ways of appraising those situations. The behavioral differences in how Observers appraise from our large-scale observational results support the lab study of Stellar et al. (2020) who found that Observers vary the themes of their responses based on the type of suffering described by the Target.

6 Do Observers Align?

Targets experience responses as highly empathetic when an observer appraises the situation in the same way (Vyskocilova et al., 2011; Watson, 2016), which requires that their appraisals align with those of the Target. Given the behavioral similarities seen between Targets and Observers in which appraisals they use, here we test whether the responses align.

Setup We calculate the probability that an Observer O's appraisal of type a_j is aligned the each type a_j when used by the Target T: $p(a_i^O|a_j^T)$.

Results In aggregate, Observers only partially aligned with how the Targets appraised (experienced) their situation (Figure 3). Instead of having the same appraisal, the majority of the Observer's aligned text was giving advice to the Target about a particular aspect.

Giving advice is a well-known aspect of Reddit support communities (e.g., De Choudhury and De, 2014) and some individuals so seek out communities for such advice (e.g., Sowles et al., 2017; O'Neill et al., 2018). While advice is not considered a component of empathy—and in some circumstances is considered counter-productive when used in empathetic situations like counseling (Barkham and Shapiro, 1986; Lieberman III and Stuart, 1999)—its frequency does highlight its importance in the lived experience of support-seeking individuals. Indeed, Depow et al. (2021) note that the experience of empathy in everyday situations encompasses a much broader set of behaviors than those listed in academic definitions.

Nevertheless, we do see a strong diagonal trend in Figure 3 that suggests that, when not giving advice, Observers do frequently align in their appraisals, suggesting empathetic behavior. Two off-diagonal trends also emerge. First, Observers frequently respond with Certainty; in our annotation, we found that these were frequently gestures meant to reassure the Target of their choice or action, and, thus, may be viewed as a type of compassionate response. Second, Observers often respond to a Target describing the objective experience with comments about the Target's agency (or not) in the situation; here too we find a type of compassion-based response where Observers use agency language to deflect responsibility from the Target onto other parties mentioned in the Target's experience. Appendix F reports additional details on specific subreddits' differences and behaviors.

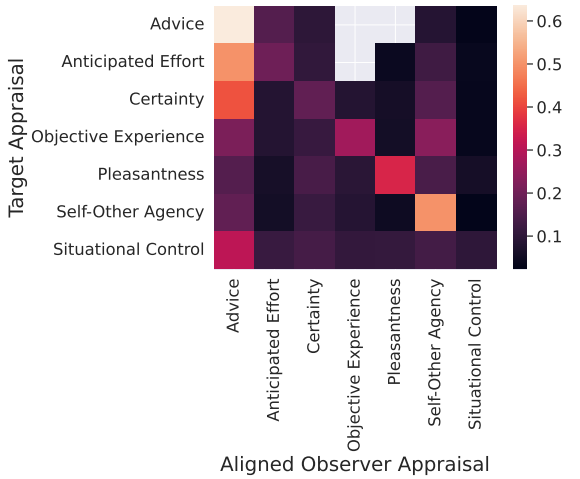

Figure 3: The probabilities of a Target's appraisal type (row) having the specific appraisal (col) in the aligned span of the Observer's. Empty cells indicate the aligned pairs are rare or non-existent in the data.

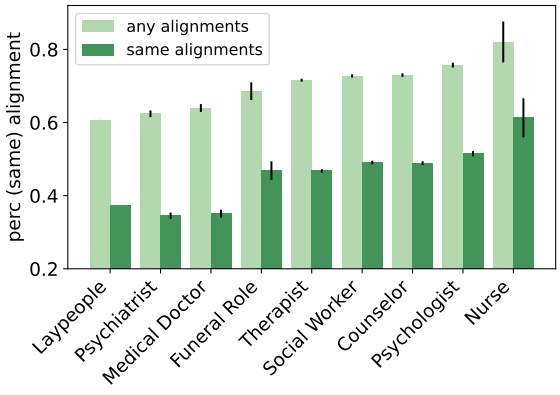

Figure 4: Mean alignment by profession; the error bar for laypeople is too small to be seen.

7 Alignment by Professional Observers

Subreddit communities contain a mixture of Observers, some of whom have professional training in mental health or medical domains. Such training frequently includes discussions of empathetic and patient-centered dialog (Hojat, 2016; Lam et al., 2011). Some communities allow users to include a flair next to their username to indicate a self-reported qualification, such as a PhD in Psychiatry. Given that users with such flairs should have experienced some training on how to behave more empathically—i.e., more alignment—here, we test the level of alignment of different professions of Observers relative to the general public in our data.

Setup We used the flair on Reddit as an indicator of whether a user is a professional or not. We adopt the flair-profession categorization of Lahnala et al. (2021) to map the text of each flair to a specific profession:[1] Counselor, Funeral Role, Medical Doctor, Nurse, Psychiatrist, Psychologist, Psychotherapist, Social Worker, or Therapist; see Appendix E for details. Flairs not indicating a specific degree or known license were left unmapped. For those with professional degrees, we also extract any reported status in training as either Fully Licensed or a Student. Professionals are defined as those who are licensed and have a non-student title, while laypeople are authors who do not have any student or professional flair throughout their usage of Reddit. We further filtered the replies from the professionals when they were of their highest/most recent training level. Ultimately, we identified 14,648 users as professionals, 2978 as students, and 1,686,362 laypeople in our data for analysis. Appendix Table 9 shows the count by profession.

Alignment is measured as the percentage of appraisals in the Target's message that have an alignment in the Observer's. We report the mean percentage for laypeople and each profession's users, calculated over all data in our dataset.

Results Mental health professionals have higher alignment than laypeople, shown in Figure 4, both for aligning with the same appraisals and in general. A small split can be seen within professions as well: Professions for clinical therapy (Therapist, Social Worker, Counselor, and Psychologist) have among the highest alignment with Targets, while medical Professions (Psychiatrists and Doctors) are much lower—even lower than laypeople at matching the same alignment. Our results suggest that the training received by mental health professionals does lead to higher alignment than laypeople.

Does the flair itself drive behavior? Individuals who list their professional degrees as flair in a subreddit often interact with others in different subreddits where no such flair is visible. With no explicit mention of their profession, there is less reputational harm in responding with less effort which could lead to lower alignment. To test whether flair visibility drives alignment, we fit a linear regression on the Observer's percent of alignment with categorical factors for (i) the profession, (ii) the subreddit, and (iii) profession flair visibility.

Flair visibility leads to a small but significant

[1]We note that some professions overlap in their theme. For example, Psychiatrists, Psychologists, Psychotherapists, and Social Workers may all be considered Therapists or Counselors. However, some qualifications, such as a Licensed Professional Counselor (LPC) do have an associated degree.

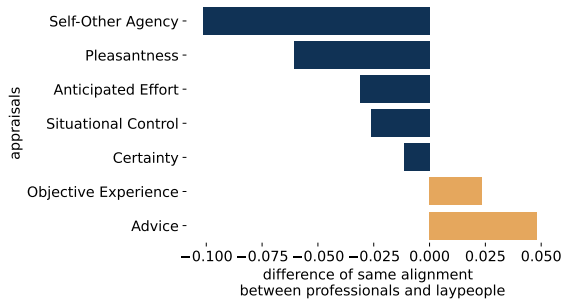

Figure 5: Difference of mean alignment between professionals and laypeople by appraisal; bars pointing right are appraisals more commonly aligned by professionals, left for laypeople.

increase in alignment (β=0.027; p<0.01); full regression results are in Appendix Table 10. However, the magnitude of the increase suggests that professionals still reply with relatively high empathy when not publicly sharing their profession. Note that this effect is also not due to the change in community, as the subreddit regression factor controls for relative differences in alignment between communities.

How do professionals differ in alignment? Given that mental health professionals better align with Targets, we examine how they differ from laypeople in which appraisals are aligned. Here, we restrict professionals to the most aligned: Therapist, Social Worker, Nurse, Psychologist, and Counselor as professionals in this analysis, while laypeople were defined as before. To control for potential differences in the content Observers are responding to, we only examine Target messages that have replies by at least one professional and one layperson. Replies were grouped by professionals and laypeople, with the percentage of the same appraisal alignment over each Target appraisal calculated. To compare professionals and laypeople, we calculate the difference in mean probability of using the same appraisals as the Target.

Surprisingly, while professionals have higher total alignment, they are much less likely to use the same appraisals in their response (Figure 5). Controlling for Target, professionals are much less likely to respond to Target's appraisals about their agency or the situation's pleasantness with the same appraisal, compared with laypeople. Instead, we find their aligned responses are more commonly advice to the Target.

8 Does experience influence alignment?

Professional health practitioner training emphasizes the importance of empathetic communication. However, multiple studies have noted that this training period marks a high point, and doctors and nurses become less empathetic over time (Wilson et al., 2012; Chen et al., 2007). Our first question is whether we observe a similar drop-off after therapist students transition to their fully licensed roles in our longitudinal data.

Laypersons too may benefit from explicit feedback on what comments are considered helpful, likely learning how to engage more empathetically over time. Redditors are known to offer such feedback when the response matches their support-seeking goals (Peng et al., 2021). As a second related question, we test whether laypersons become more empathetic as they receive such feedback on which comments were most helpful.

Setup For the first question, we collect all authors who have flairs with licensed or student training levels. We then collected all of their engaged conversations as observers and split their comments into licensed and student periods based on when the flair text changes. Comparison is made between conversations involving licensed observers and student observers. For the second question, we use data from *r/Advice*, which assigns flairs of different levels to users based on how many times they have replied as an Observer and another user has replied to express gratitude for their comment. We treat these flairs as proxies for experience in writing helpful replies. We collected the replies from different experience levels (flairs) and calculated the mean percentage of alignment for each level.

Results Our results show that students are often more empathetic than their fully licensed counterparts (Figure 6). Of these, only the drop for Therapist users is significant at p<0.1 using an independent t-test. Our observations mirror results from Wilson et al. (2012) and Chen et al. (2007) showing that nursing and medical students decreased their empathy levels with patients as they received more training. One likely driver of such drop-off is compassion fatigue, where high levels of empathy towards Targets in a therapeutic setting can lead to a decreased ability to feel compassion for others (Turgoose and Maddox, 2017).

A similar trend is seen for users in r/Advice as they gain more experience making helpful comments (Figure 7), where users initially comment

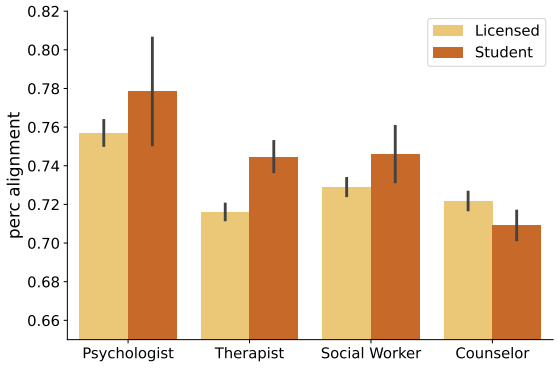

Figure 6: Mean alignment comparison between the licensed and students

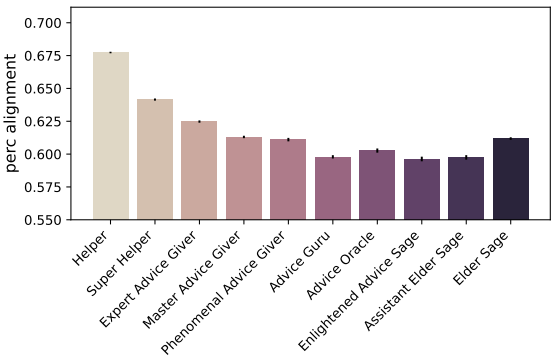

Figure 7: Mean alignment by the amount of gratitude received at the time of comment, shown as flair from r/Advice ordered from least to most thanked.

with very high alignment and then slowly drop in alignment during their continued engagement until they reach a state-state level. While status-seeking is a known strong motivator for Reddit users (Moore and Chuang, 2017), we hypothesize that as users engage more frequently, some drop off in empathetic alignment may come from social media fatigue, which can decrease psychological well-being (Dhir et al., 2018) and thus lead to a decreased ability to empathize. We also hypothesize, that similar to professionals, these users also experience compassion fatigue from engaging with distressing comments.

9 Discussion

Individuals seek support in online communities, yet the type of support they receive varies, and, as we show, may not be well-aligned or, when aligned, may be advice rather than validation of their experience. The result that healthcare professionals are more likely to give advice raises new research questions about how online and offline therapeutic practices may differ. Future studies could examine (i) the qualitative differences in the advice of laypersons vs professionals, (ii) whether Targets view this advice as more valuable or empathetic, and (iii) if possible, the underlying motivations for professionals to give more advice, despite their training. The models developed in this paper can help surface such examples for study.

The task of empathetic alignment can provide new opportunities for advancing NLP. For example, generating empathetic responses is effortful for many people (Cameron et al., 2019), in part, because of the mental work. NLP models for recognizing this alignment could be used for assistive technologies that lower the cognitive load for responding with high empathy, such as highlighting passages that an Observer might respond to and assessing whether their responses match what a Target has actually said. Such tools could potentially provide lower-effort entry points into the conversation to help people engage.

10 Conclusion

Empathy requires perspective taking on the part of an Observer to align their cognitive and emotional experiences with another. This study goes beyond prior work in NLP on empathy to make these empathetic alignments explicit and to identify how observers mirror (or miss) the types of perspectives described by Targets. By developing a new dataset, ALOE, and models for appraisals and alignments in empathetic dialogues, our work enables studies of how and when Observers empathize. In a large-scale study of Reddit, we show that individuals seeking mental health support do receive empathetic replies—but that many aligned responses are giving advice, rather than acknowledging their perspective. However, we also show that mental health professionals on Reddit show much higher alignment than the general public. Our data and model can support future studies on how to help identify and correct misalignment when drafting responses or suggest opportunities for new alignments. All data, models, and annotation materials are available at https://github.com/jessicayjm/modeling_empathy_alignment and the annotation tool is available in a stand-alone form at https://github.com/jessicayjm/span_alignment_annotation_tool

11 Limitations

Our study examines conversations on a public social media platform, Reddit. Thus, our results may not be generalizable to other settings such as in-person settings where modalities other than texts might also be significant, or where individuals can speak in full confidence of anonymity. However, having the benefits of using longer texts, we were able to perform analysis on more complex appraisals compared with methods using shorter texts such as Tweet data or text messages.

Secondly, we only analyzed Reddit posts and the top-level comments and replies to those posts. While these post-comment pairs offer the cleanest signal of individuals looking for mental health support, our focus necessarily limits empathetic conversation that may be happening in replies to comments. We view this as an opportunity for future work in multi-turn dialog in Reddit. Meanwhile, analysis from the Reddit data still carries its significance for longer conversation analysis given the impressive number of users being active on the platform.

A key limitation in annotation comes from the inaccessibility to the mindsets of original targets and observers. Our dataset reflects third-party perceptions of the state of mind of annotators—a challenging task given that we lack information on who the targets and observers are. However, under the design of the annotation process, our annotations likely mirrored the process that the observers go through when assessing potential targets. Further, our adjudication process ensured that multiple potential interpretations were considered when any message was ambiguous so that the most likely could be chosen. Future work should be encouraged to collect first-hand annotations directly from targets and observers, as this is currently a missing dataset for the community.

We should also be aware of the inherent biases of lived experience from annotators. Though crowdsourcing could provide more diversity, the annotation task itself is complex and not immediately amenable to crowdwork without extensive validation. We hope that our work can set a baseline for further generalization and to test how annotator identity can influence perceptions of empathy and alignment.

The general trends in our experiments rely on classifiers that imperfectly learn how to identify appraisals and how Targets and Observers align. Given the challenge of these classification tasks, our initial models attain only moderate performance which could potentially influence our downstream results. As a result, we have taken care to only describe trends in aggregate and to report confidence intervals and standard errors wherever possible. While moderate in performance, our approach mirrors work in other NLP tasks such as framing (e.g., Ajjour et al., 2019; Akyürek et al., 2020; van den Berg and Markert, 2020; Dayanık et al., 2022; Mendelsohn et al., 2021) where models operate on nuanced, often social, data to identify a moderate number of labels in order to derive large-scale trends when applying these classifiers at scale. Nonetheless, social information remains challenging to recognize even for large language models (Choi et al., 2023) and our dataset provides an opportunity for future work to improve performance at recognizing empathetic alignment, which can open new doors for more fine-grained analyses of empathetic behavior.

Last, our results are drawn from a primarily Western social media context and a Western-educated annotation pool. This cultural backdrop likely limits the generalizability of our results to other cultures. In our work, the annotators were aware of the cultural context of the paper's data. Other work will be needed to understand how individuals from a variety of cultural backgrounds appraise distressing settings and how they effectively engage with empathy.

12 Ethical Considerations

The work includes some comments by people who have experienced or are experiencing distressing events. While this data is fully public—posted by the authors themselves publicly to seek support—additional views of the data could risk having them being further re-exposed to the events by malicious actors. However, we view this risk as very low compared with the potential benefits of studying distressing events by providing insights for helping Observers better engage with Targets, which would lead to long-term support in the community.

Considering the data source is public but sensitive, we only release the data to researchers after filling out a request that acknowledges the potentially-sensitive nature of the data and the responsibility of its use.

References

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. 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 2922–2932.

Afra Feyza Akyürek, Lei Guo, Randa Elanwar, Prakash Ishwar, Margrit Betke, and Derry Tanti Wijaya. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8614–8624.

Michael Barkham and David A Shapiro. 1986. Counselor verbal response modes and experienced empathy. *Journal of Counseling Psychology*, 33(1):3.

C Daniel Batson, Bruce D Duncan, Paula Ackerman, Terese Buckley, and Kimberly Birch. 1981. Is empathic emotion a source of altruistic motivation? *Journal of personality and Social Psychology*, 40(2):290.

R.J.R. Blair. 2005. Responding to the emotions of others: Dissociating forms of empathy through the study of typical and psychiatric populations. *Consciousness and Cognition*, 14(4):698–718. The Brain and Its Self.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, page 737–744, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

C Daryl Cameron, Cendri A Hutcherson, Amanda M Ferguson, Julian A Scheffer, Eliana Hadjiandreou, and Michael Inzlicht. 2019. Empathy is hard work: People choose to avoid empathy because of its cognitive costs. *Journal of Experimental Psychology: General*, 148(6):962.

Daniel Chen, Robert Lew, Warren Hershman, and Jay Orlander. 2007. A cross-sectional measurement of medical student empathy. *Journal of general internal medicine*, 22:1434–1438.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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.

Mark H Davis. 2018. *Empathy: A social psychological approach*. Routledge.

Erenay Dayanık, Andre Blessing, Nico Blokker, Sebastian Haunss, Jonas Kuhn, Gabriella Lapesa, and Sebastian Pado. 2022. Improving neural political statement classification with class hierarchical information. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2367–2382.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80.

Gregory John Depow, Zoë Francis, and Michael Inzlicht. 2021. The experience of empathy in everyday life. *Psychological Science*, 32(8):1198–1213.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Amandeep Dhir, Yossiri Yossatorn, Puneet Kaur, and Sufen Chen. 2018. Online social media fatigue and psychological wellbeing—a study of compulsive use, fear of missing out, fatigue, anxiety and depression. *International Journal of Information Management*, 40:141–152.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. 2021. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*.

Nancy Eisenberg, Tracy L Spinrad, and Amanda S Morris. 2013. Prosocial development.

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *Proceedings of the third workshop on computational linguistics and clinical psychology*, pages 63–73.

Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction.

Judith A Hall, Rachel Schwartz, and Fred Duong. 2021a. How do laypeople define empathy? *The Journal of social psychology*, 161(1):5–24.

Judith A Hall, Rachel Schwartz, Fred Duong, Yuan Niu, Manisha Dubey, David DeSteno, and Justin J Sanders. 2021b. What is clinical empathy? perspectives of community members, university students, cancer patients, and physicians. patient education and counseling. *Patient education and counseling*, 104(5):1237—-1245.

Terry Hanley, Julie Prescott, and Katalin Ujhelyi Gomez. 2019. A systematic review exploring how young people use online forums for support around mental health issues. *Journal of mental health*, 28(5):566–576.

Elaine Hatfield, Richard L Rapson, and Yen-Chi L Le. 2011. Emotional contagion and empathy. *The social neuroscience of empathy.*, page 19.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Meghan L Healey and Murray Grossman. 2018. Cognitive and affective perspective-taking: evidence for shared and dissociable anatomical substrates. *Frontiers in neurology*, 9:491.

Martin L Hoffman. 2001. *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

Mohammadreza Hojat. 2016. Empathy in health professions education and patient care.

Mohammadreza Hojat, Daniel Z Louis, Vittorio Maio, and Joseph S Gonnella. 2013. Empathy and health care quality.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans.

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.

Tony Chiu Ming Lam, Klodiana Kolomitro, and Flanny C Alamparambil. 2011. Empathy training: Methods, evaluation practices, and validity. *Journal of Multidisciplinary Evaluation*, 7(16):162–200.

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):42–58.

Joseph A Lieberman III and Marian R Stuart. 1999. The bathe method: incorporating counseling and psychotherapy into the everyday management of patients. *Primary care companion to the Journal of clinical psychiatry*, 1(2):35.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Alvin R. Mahrer. 1997. Empathy as therapist-client alignment.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263.

Carrie Moore and Lisa Chuang. 2017. Redditors revealed: Motivational factors of the reddit community. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Maria Moudatsou, Areti Stavropoulou, Anastas Philalithis, and Sofia Koukouli. 2020. The role of empathy in health and social care professionals. In *Healthcare*, volume 8, page 26. MDPI.

John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences*, 25(2):113–122.

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

Tully O'Neill et al. 2018. 'today i speak': Exploring how victim-survivors use reddit. *International journal for crime, justice and social democracy*, 7(1):44–59.

Zhenhui Peng, Xiaojuan Ma, Diyi Yang, Ka Wing Tsang, and Qingyu Guo. 2021. Effects of support-seekers' community knowledge on their expressed satisfaction with the received comments in mental health communities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Kelley Raab. 2014. Mindfulness, self-compassion, and empathy among health care professionals: a review of the literature. *Journal of health care chaplaincy*, 20(3):95–108.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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

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.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

A. Smith, A. Sen, and R.P. Hanley. 2010. *The Theory of Moral Sentiments*. Penguin Classics. Penguin Publishing Group.

Adam Smith. 2006. Cognitive empathy and emotional empathy in human behavior and evolution. *The Psychological Record*, 56(1):3–21.

Shaina J Sowles, Melissa J Krauss, Lewam Gebremedhn, and Patricia A Cavazos-Rehg. 2017. "i feel like i've hit the bottom and have no idea what to do": Supportive social networking on reddit for individuals with a desire to quit cannabis use. *Substance abuse*, 38(4):477–482.

Jennifer E Stellar, Craig L Anderson, and Arasteh Gatchpazian. 2020. Profiles in empathy: Different empathic responses to emotional and physical suffering. *Journal of Experimental Psychology: General*, 149(7):1398.

Jennifer E. Stellar and Fred Duong. 2023. The little black box: Contextualizing empathy. *Current Directions in Psychological Science*, 32(2):111–117.

Richard Thwaites and James Bennett-Levy. 2007. Conceptualizing empathy in cognitive behaviour therapy: Making the implicit explicit. *Behavioural and Cognitive Psychotherapy*, 35(5):591–612.

S Kay Toombs. 2001. The role of empathy in clinical practice. *Journal of Consciousness Studies*, 8(5-6):247–258.

David Turgoose and Lucy Maddox. 2017. Predictors of compassion fatigue in mental health professionals: A narrative review. *Traumatology*, 23(2):172.

Esther van den Berg and Katja Markert. 2020. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6315–6326.

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

Jana Vyskocilova, Jan Prasko, and Milos Slepecky. 2011. Empathy in cognitive behavioral therapy and supervision. *Activitas Nervosa Superior Rediviva*, 53(2):72–83.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Jeanne C Watson. 2016. The role of empathy in psychotherapy: Theory, research, and practice. In *Humanistic psychotherapies: Handbook of research and practice (2nd Edition)*, pages 115–145. American Psychological Association.

Anuradha Welivita, Chun-Hung Yeh, and Pearl Pu. 2023. Empathetic response generation for distress support. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 632–644.

Sarah E Wilson, Julie Prescott, and Gordon Becket. 2012. Empathy levels in first-and third-year students in health and non-health disciplines. *American journal of pharmaceutical education*, 76(2):24.

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

Joshua Wondra and Phoebe Ellsworth. 2015. An appraisal theory of empathy and other vicarious emotional experiences. *Psychological review*, 122.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. *arXiv preprint arXiv:2105.06829*.

Chengkun Zeng, Guanyi Chen, Chenghua Lin, Ruizhe Li, and Zhigang Chen. 2021. Affective decoding for empathetic response generation.

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.

Ling Yu Zhu, Zhengkun Zhang, Jun Wang, Hongbin Wang, Haiying Wu, and Zhenglu Yang. 2022. Multi-party empathetic dialogue generation: A new task for dialog systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 298–307.

A Subreddits Used for Data

A.1 Subreddits for building alignment dataset

anxiety, depression, Miscarriage, domesticviolence, widowers, GriefSupport, Petloss, FiftyFifty, SuicideBereavement, ttcafterloss, heart-

break, BreakUps, BreakUp, BipolarSOs, dementia, Alzheimers, ExNoContact, CautiousBB, domesticviolence, CaregiverSupport, abusiverelationships, emotionalabuse, marriageadvice, lastimages, PrayerRequests, OldManDog, seniorkitties, askfuneraldirectors, death, dogpictures, MadeMeCry, cancer, MomForAMinute, sad, happycryingdads

A.2 Subreddits for Reddit tree

depression, BPD, dementia, Vent, abusiverelationships, offmychest, lonely, BreakUps, SOCIALSKILLS, CautiousBB, Advice, TalkTherapy, MomForAMinute, adultsurvivors, getdisciplined, MadeMeCry, NarcissisticAbuse, bipolar, SuicideWatch, Anxiety, widowers, selfharm, GriefSupport, BPDlovedones, SingleParents, Anger, mentalhealth, datingoverforty, heartbreak, emotionalabuse, ExNoContact, lastimages, PrayerRequests, PregnancyAfterLoss, marriageadvice, DecidingToBeBetter, SuicideBereavement, CPTSD, socialanxiety, seniorkitties, IWantToLearn, OldManDog, Petloss, ttcafterloss, cancer, psychotherapy, OCD, datingoverfifty, emotionalneglect, Alzheimers, BorderlinePDisorder, Codependency, self-improvement, death, gaslighting, BPDPartners, productivity, dbtselfhelp, CaregiverSupport, NarcAbuseAndDivorce, MMFB, therapy, Miscarriage, domesticviolence, BipolarSOs, BreakUp, askatherapist, sad, LifeAfterNarcissism, AskPsychiatry, FriendsOver40, NarcissisticSpouses, ChildrenofDeadParents, BPDlite, happycryingdads, CBT, narcissism, Grieving, BodyAcceptance, MentalHealthUK, BPD4BPD, askfuneraldirectors, InternalFamilySystems, CPTSDNextSteps, EmotionalAbuseSupport, CancerCaregivers, cptsdcreativeas, BPDrecovery, grief, traumatoolbox, COVIDgrief

B Pre-Annotated Data Filtering

We used all three models (Distress, Condolence, and Empathy classifiers) from Zhou and Jurgens (2020) to filter the Reddit data. Both distress and condolence classifiers are bert-base-uncased models, while the empathy classifier is a roberta-base model. Filtering was performed to surface data likely to contain empathy. We first identify all posts where p(distress)>0.9, then we retain comment replies rated as p(condolence)>0.9. From these post-comment pairs, we retain all pairs with an empathy rating of at least 2 on their 5-point scale.

C Additional Annotation Details

C.1 Annotation Instructions

Annotators were instructed to read the 11-page annotation codebook (included in the Supplementary Data). The codebook contains detailed instructions and examples for each appraisal type and the three new span types we introduce. Table 4 shows the general definition in the codebook for each.

C.2 Annotator Recruitment

Annotators were recruited from a large mailing list of university undergraduates. Interested undergraduates participated in an initial paid one-hour training session and five (4 women, 1 man) signed on to continue annotating after an initial vetting of their work done during the training session. Annotators were paid $15/hr USD and were able to work up to 10 hours per week, with flexibility depending on their schedule.

C.3 Annotation Website Interface

Annotators used a custom website for all their work. The interface for appraisal annotation is shown in Figure 8. Annotators can select labels and highlight spans for annotation, track the annotation process, and make private notes.

The interface for alignment annotation is shown in Figure 9. Annotators can click one span from Target and one span from Observer to annotate alignments. The interface shares the note function with the appraisal annotation interface.

The review interface for both appraisal and alignment is the same as the annotation interface but with an extra discussion function where annotators can post their comments, raise questions regarding each instance, and communicate with each other asynchronously.

The interfaces for the admin user to finalize the appraisal annotations are shown in Figure 10 and Figure 11. The admin has access to all annotators' work and is able to decide the final annotation. The discussion panel shows all public comments from annotators.

Appraisal/Span	General Definition
Pleasantness	How pleasant the situation was.
Anticipated Effort	How much effort was needed to deal with the situation.
Situational Control	How much the situation was out of anyone's control.
Self-other Agency	How much oneself or another person was responsible for the situation.
Attentional Activity	Reflects how much the Target's attention was drawn to the situation rather than diverted away from the situation. This appraisal has more to do with the occurrence of an event—its suddenness, familiarity, and predictability—rather than the qualities of the event like its pleasantness.
Certainty	Certainty about what was happening in the situation or what would happen next
Objective Experience	Description of the experience of the author that is not an appraisal or the broader context/circumstances in which their story takes place.
Advice	Expressions of asking or providing advice.
Trope	General sympathetic expressions that are not specific to the Target.

Table 4: General definitions used in the annotation codebook for each of the appraisal types. In the codebook, each definition is followed by more details, notes, and examples of what is or is not that appraisal for both observers and targets.

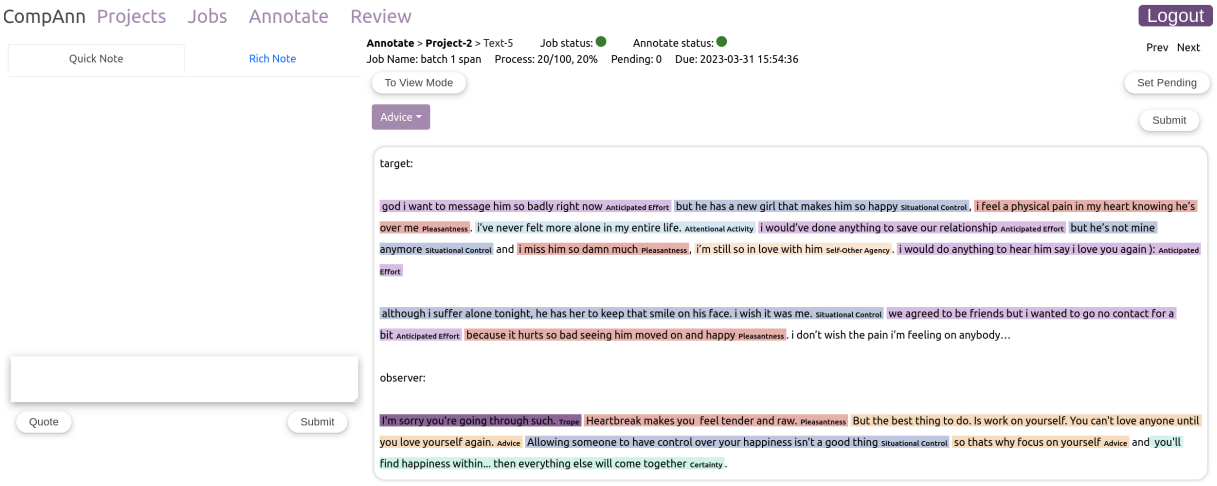

Figure 8: Appraisal Annotation interface.

The interfaces for the admin user to finalize the alignment annotations are shown in Figure 12. The admin can directly select alignment from the annotators' work. New alignments that are not identified by annotators can be added through *Finalize Alignment* panel which is the same as the alignment annotation interface.

C.4 Additional Annotated Examples

Table 6 shows additional examples of appraisals and other categories that were annotated.

C.5 Addition Observations on Labeling

Attentional Activity was rare in our data, in part, because the general perception was that other types of appraisals were more salient and likely explanations. For example, a strongly *Attentional Activity* dominated span could be: "On the one hand I don't want to go around starting every conversation announcing that my brother has passed, but it's been THE central event in my life recently and the biggest thing on my mind." However, in many cases, other appraisals will dominate the interpreta-

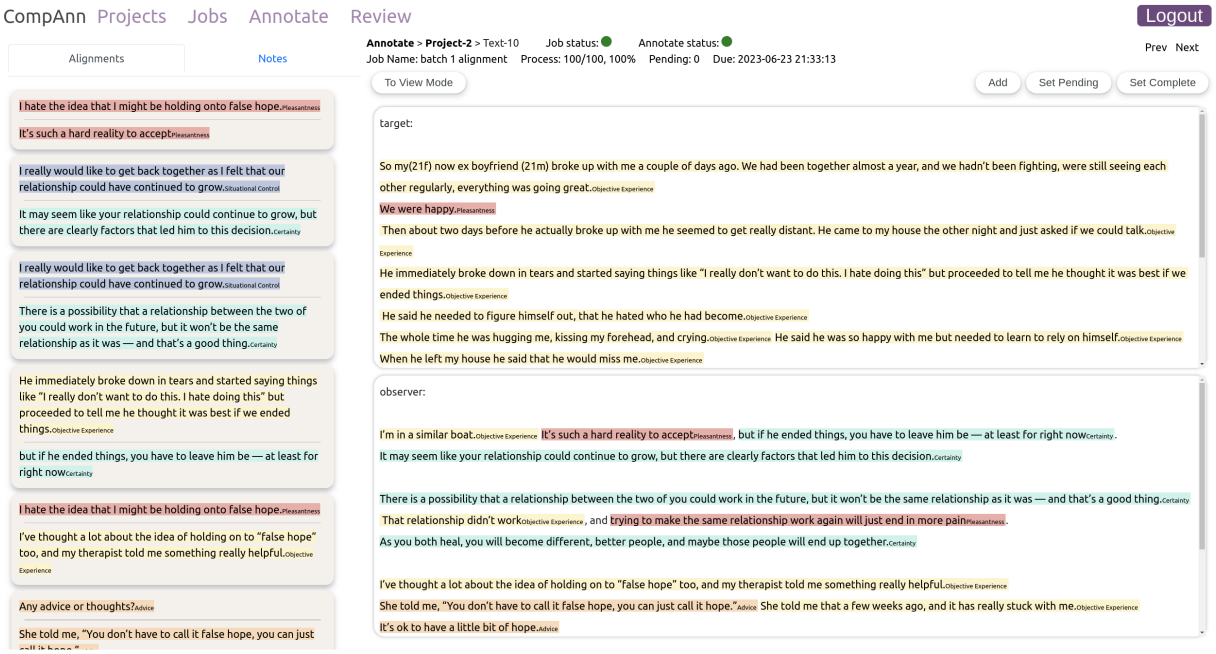

Figure 9: Alignment Annotation interface.

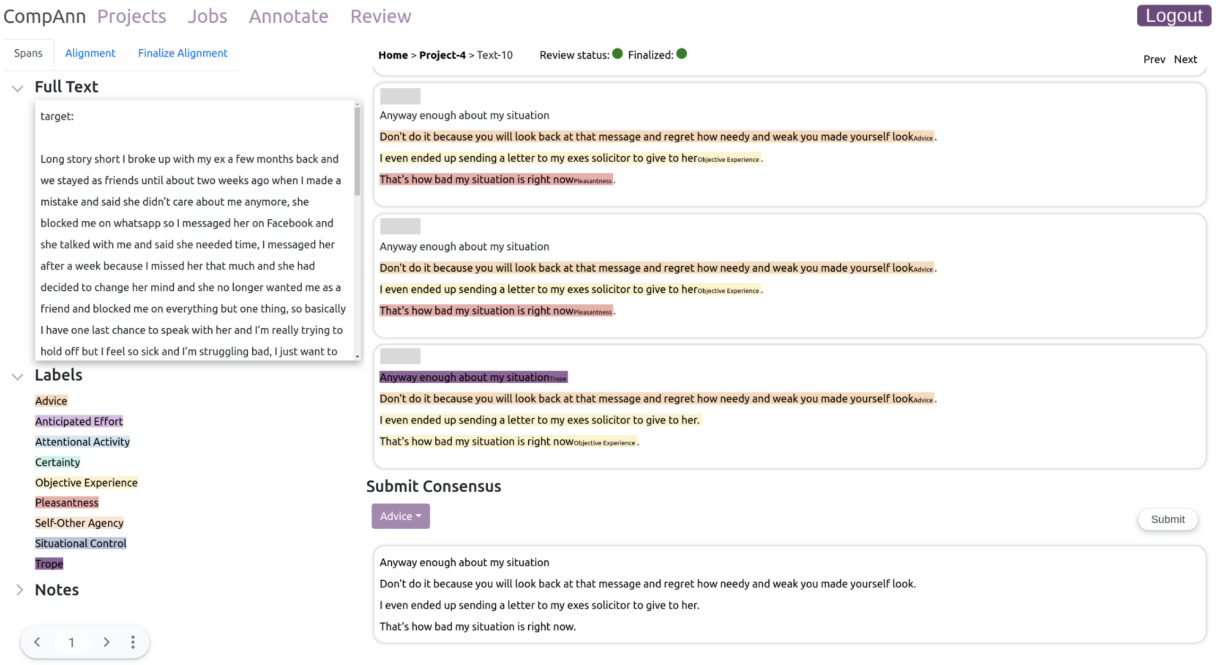

Figure 10: Finalize appraisal annotation interface.

tion, such as in the following examples:

- *Pleasantness* dominates: "I've never felt more alone in my entire life."
- *Anticipated Effort* dominates: "Depression was and is still the hardest challenge that I face everyday."
- *Objective Experience* dominates: "Called her this morning and police picked up saying she is dead."

D Additional Model Details and Results

D.1 Training Details

Information on the size of the different dataset splits is shown in Table 5, as well as what percent of the data was originally labeled with multiple

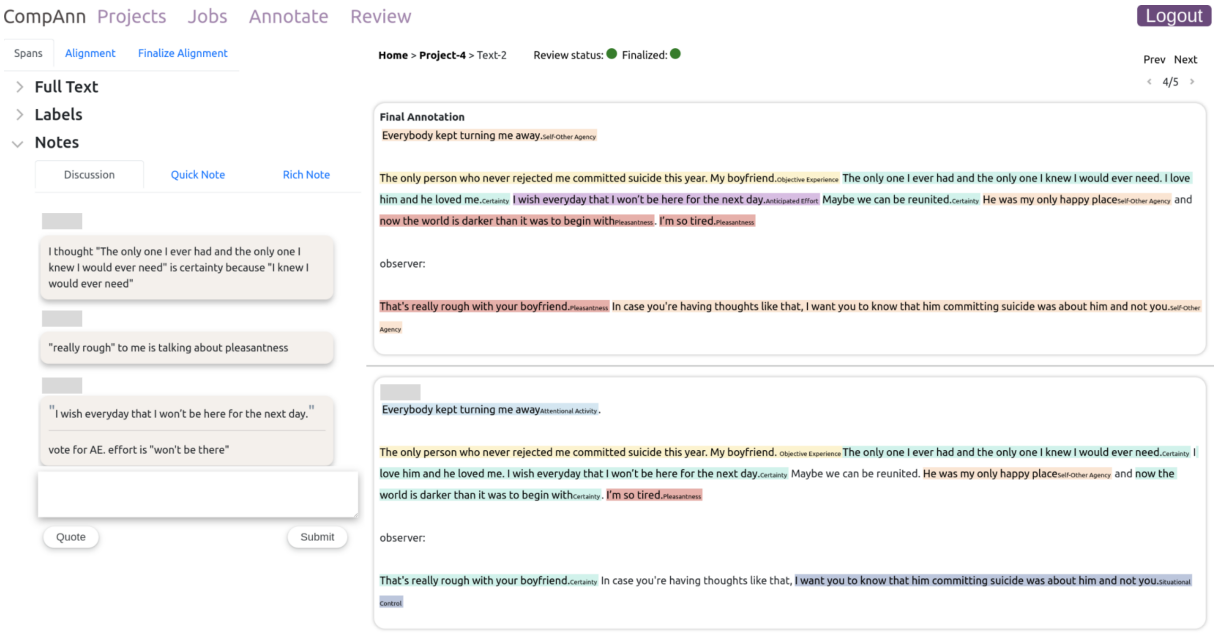

Figure 11: Finalize alignment annotation interface with note function.

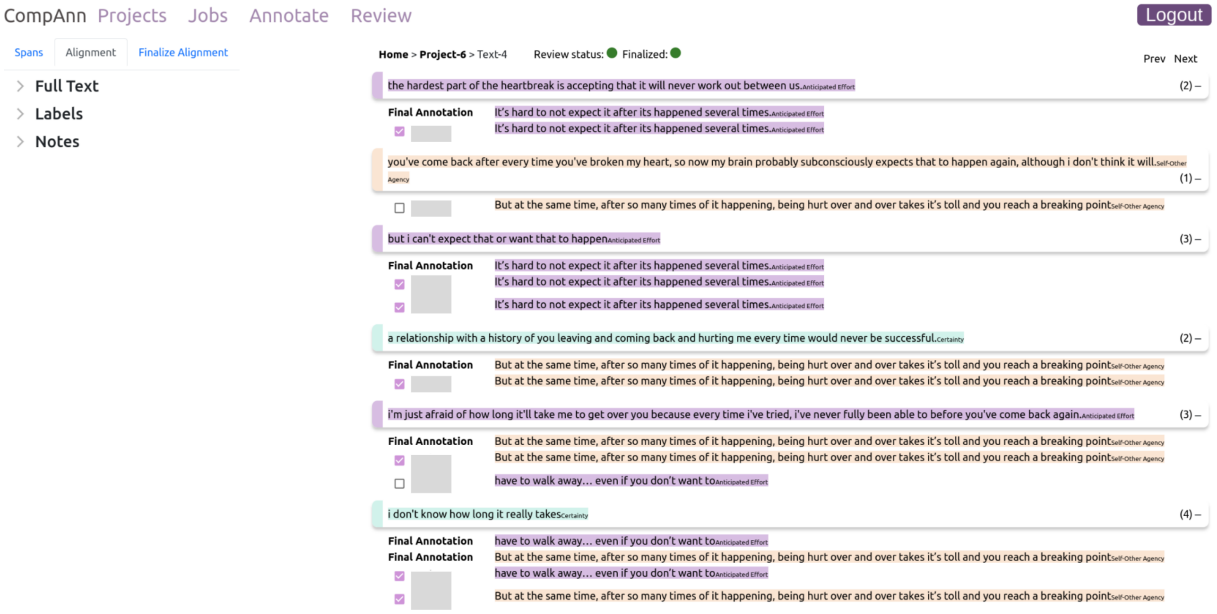

Figure 12: Finalize alignment annotation interface: view and select from annotators.

spans of different appraisal types (11%).

All parameters not mentioned use default values in Huggingface transformers library (Wolf et al., 2020). The random seed is set to be 0 for all the training.

D.2 Span prediction

All span prediction models are trained on cross-entropy loss with AdamW optimizer (Loshchilov and Hutter, 2019). OpenPrompt+RoBERTa-large is trained with a learning rate of 1e-7, while all other models are trained with a learning rate of 1e-6. Specifically for OpenPrompt models: `freeze_lm=False`, `max_seq_len=512`,

	Train	Dev	Test
No Label	461	49	133
Pleasantness	956	124	229
Anticipated Effort	767	113	198
Certainty	934	163	281
Objective Experience	1146	144	357
Self-Other Agency	1161	252	369
Situational Control	782	112	229
Advice	1120	181	321
Trope	262	31	80
Total	7589	1169	2197
Multi-label	872	130	234

Table 5: Number of sentences for appraisal prediction containing both Target and Observer. Multi-label shows how many sentences contain more than one appraisal.

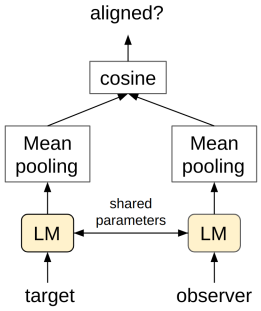

Figure 13: Alignment model architecture.

`decoder_max_len=3`, `teacher_forcing=False`, `truncation_method=head`. All other specific information on the training process and hyperparameters are shown in Table 7.

D.3 Alignment prediction

The model architecture is shown in Figure 13. Both all-MiniLM-L6-v2 and all-mpnet-base-v2 are trained on mean squared error loss (mean reduction) with AdamW optimizer, `max_epoch=300`, `patience=15`, `batch_size=16`. The prediction threshold is set to be 0.3 ($p \geq 0.3$ will be predicted as aligned). all-MiniLM-L6-v2 reached the lowest dev loss at epoch 4 with a total training time of 1 hour and 15 minutes. all-mpnet-base-v2 reached the lowest dev loss at epoch 9 with a total training time of 5 hours and 7 minutes.

D.4 Model Performance

Detailed model performances for appraisal prediction are shown in Table 3.

E Additional Profession Results

The full breakdown of user and comment counts for those users with profession flairs is shown in Table 9.

Table 10 shows the full linear regression model fit for predicting the level of alignment based on a user's profession, the subreddit, and whether their profession's flair was visible at the time of comment.

F Subreddit-level Differences and Analysis

While we report aggregate statistics and trends, the subreddits in our studies still constitute distinct communities that vary in their behaviors (cf. Figures 1 and 2). Following, we report on two differences between subreddits that underscore themes in the main paper: the prevalence of giving Advice and the misalignments between Targets and Observers.

F.1 Which subreddits give more Advice when asked?

Setup We calculated the percentage of alignments where the Observer responded with Advice (i.e., the Target has requested advice), and averaged by subreddits.

Results While advice is present in all subreddits, as Figure 14 shows, not all communities give advice. Advice appears the most frequently when the Targets actively ask for it (cf. Figure 3), yet in topics of loss and grief, Advice becomes much less frequent in the conversation despite the request for advice. However, at the other extreme, Targets in subreddits related to mental health receive an abundance of advice. We hypothesize that in loss and grief subreddits, Observers instead focus on emotional support rather than suggestions, in part because of the difficulty of identifying what specifically can be done in such circumstances.

F.2 Misalignment for Appraisals

Given the topical differences, do Observers in some subreddits align more closely with their Targets?

Setup For each appraisal/span in Target, we computed the percentage where the aligned appraisal in Observer is different from the Target and averaged by subreddits.

Appraisals
(*Pleasantness*) I feel so absolutely shattered into infinite pieces and even that doesn't seem to express this enough.
(*Anticipated Effort*) I know it's important to live in the now but right now I'd rather sleep than face the reality of anything.
(*Situational Control*)
Whenever I let myself feel joy in life again my brain says "hey imagine if you were feeling this joy with your ex, remember that?" and my heart starts to hurt again.
(*Self-Other Agency*) She broke up with me, she destroyed my heart and she's the one finding someone new to be happy with while i haven't had even a crush on anyone in all this time.
(*Attentional Activity*) I never imagined it would be this hard to cope, especially for a pregnancy that ended so early.
(*Certainty*) I don't know if I ever wanted the relationship to begin with or if I just wanted validation.
Non-appraisals
(*Objective Experience*) I posted about my sweet girl Lulu just a bit over a month ago. Since then her thyroid tumor masses have gotten so big that they have started compressing her throat and now she can't eat.
(*Trope*) I'm so sorry you have to go through this.
(*Advice*) Sometimes it helps to share your good memories, either in places like this one, or with other family/friends who knew him.

Table 6: Examples on appraisals and non-appraisals.

	max epoch	**patience**	**batch size**	**best epoch**	**time**
BERT	200	15	32	23	02:35
RoBERTa	200	15	32	10	01:39
SpanBERT	200	10	32	14	01:38
DeBERTa	200	10	16	7	01:47
MiniLM	200	10	16	38	00:22
OpenPrompt+BERT	200	10	16	11	03:56
OpenPrompt+ RoBERTa	200	20	16	52	14:02
OpenPrompt+ T5-large	200	10	8	57	21:37

Table 7: Training information for predicting appraisals.

Results Table 11 shows the most and least aligned subreddits for each appraisal type. Observers align well with targets around subreddits on loss and grief in Pleasantness, Certainty, and Objective Experience. Anticipated Effort and Situational Control are mainly aligned best with topics around mental health. Noticeably, Self-other Agency has a bias towards alignment for abuse-related topics. As confirmed with what we have observed, Advice appears in observers the most when it's actively been asked for.

Less clustered than most aligned topics, appraisals could be misaligned in a diverse range of subreddits. However, the exception appears for Self-other agency, which observers seldom correctly align with targets in mental health topics.

G Model output examples on alignment prediction: qualitative error analysis

Table 12 shows examples of positive and negative classification errors for alignment prediction, along with descriptions of what pattern was seen for this type of error.

	random	majority	BERT	RoBERTa	SpanBERT	DeBERTa	MiniLM	Open-Prompt +BERT	Open-Prompt +RoBERTa	Open-Prompt+ T5-large
Macro-F1	0.11	0.03	0.38	**0.56**	0.52	0.55	0.49	0.53	**0.56**	**0.56**
Macro-Recall	0.11	0.02	0.38	0.56	0.52	0.55	0.50	0.53	**0.57**	0.56
Macro-Precision	0.11	0.11	0.41	0.57	0.54	0.56	0.51	0.55	0.58	**0.59**
Per Label Recall										
No Label	0.06	0.00	0.14	**0.34**	0.23	**0.34**	0.18	0.32	**0.34**	0.25
Pleasantness	0.11	0.00	0.51	0.66	**0.69**	0.66	0.66	0.68	**0.69**	**0.69**
Anticipated Effort	0.08	0.00	0.32	0.44	0.43	0.41	0.41	0.42	0.46	0.48
Certainty	0.12	0.00	0.36	0.48	0.54	0.44	0.50	0.46	**0.58**	0.55
Objective Experience	0.16	0.00	0.45	0.67	0.58	0.55	0.61	**0.68**	0.58	0.60
Self-Other Agency	0.15	0.18	0.42	0.62	0.62	**0.63**	0.59	0.61	0.62	0.60
Situational Control	0.10	0.00	0.11	0.31	0.21	**0.45**	0.14	0.20	0.31	0.38
Advice	0.15	0.00	0.58	**0.73**	0.69	0.70	0.70	0.71	0.72	**0.73**
Trope	0.04	0.00	0.56	0.79	0.71	0.76	0.70	0.70	**0.80**	0.79
Per Label Precision										
No Label	0.11	0.00	0.38	0.63	0.54	0.46	0.44	0.55	0.64	**0.70**
Pleasantness	0.13	0.00	0.35	0.56	0.54	**0.57**	0.47	0.51	0.54	0.52
Anticipated Effort	0.10	0.00	0.25	0.45	0.39	**0.47**	0.40	**0.47**	0.46	0.45
Certainty	0.11	0.00	0.37	**0.55**	0.49	**0.55**	0.46	0.53	0.47	0.54
Objective Experience	0.11	0.00	0.52	0.63	0.68	**0.70**	0.66	0.59	0.69	0.66
Self-Other Agency	0.11	1.00	0.38	0.53	0.51	0.53	0.50	0.50	0.55	0.54
Situational Control	0.10	0.00	0.42	0.51	0.39	0.34	0.39	0.48	**0.55**	0.52
Advice	0.11	0.00	0.44	0.63	0.62	**0.67**	0.63	0.61	0.66	0.64
Trope	0.14	0.00	0.58	0.70	0.68	0.70	0.63	0.71	0.67	0.68

Table 8: Appraisal model performance.

Profession	#Users	# Comments
Nurse	3	23
Funeral Role	21	152
Medical Doctor	24	712
Psychiatrist	17	1221
Psychologist	114	1769
Counselor	241	3374
Social Worker	338	4049
Therapist	377	4937

Table 9: Counts of how many users had valid flairs associated with each profession and the number of comments associated with each.

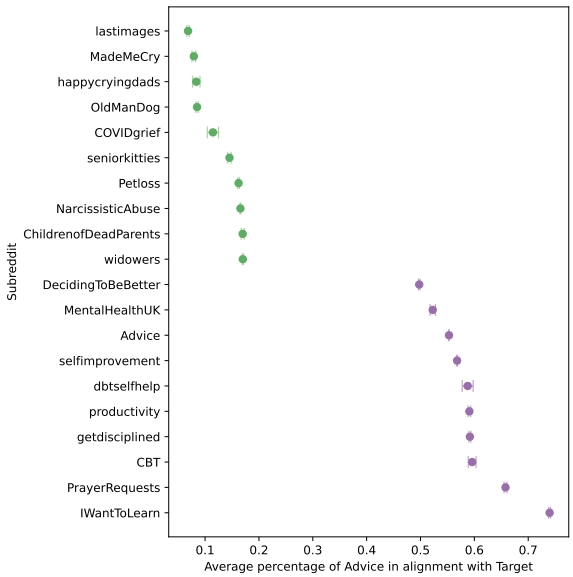

Figure 14: Percentage of alignments that Observer responding with Advice for each subreddit. Bottom-10 (least frequent) and top-10 (most frequent) are shown.

	Dependent variable:
	% Alignment
Profession: Funeral Role	0.121 (0.097)
Profession: Medical Doctor	0.028 (0.018)
Profession: Nurse	0.060 (0.058)
Profession: Psychiatrist	0.008 (0.017)
Profession: Psychologist	−0.002 (0.008)
Profession: Social Worker	0.011* (0.006)
Profession: Therapist	−0.002 (0.005)
is_title_visibleTrue	**0.027*** (0.009)**
Subreddit: adultsurvivors	0.045 (0.107)
Subreddit: Advice	0.204* (0.106)
Subreddit: Anger	−0.042 (0.295)
Subreddit: Anxiety	0.193* (0.110)
Subreddit: askatherapist	0.105 (0.105)
Subreddit: askfuneraldirectors	0.012 (0.145)
Subreddit: AskPsychiatry	0.049 (0.106)
Subreddit: bipolar	0.123 (0.106)
Subreddit: BipolarSOs	0.391 (0.295)
Subreddit: BodyAcceptance	0.129 (0.191)
Subreddit: BorderlinePDisorder	0.029 (0.173)
Subreddit: BPD	0.141 (0.109)
Subreddit: BPDlite	−0.525* (0.295)
Subreddit: BPDlovedones	0.099 (0.129)
Subreddit: BreakUp	−0.059 (0.110)
Subreddit: BreakUps	0.174 (0.111)
Subreddit: cancer	0.075 (0.114)
Subreddit: CaregiverSupport	0.047 (0.123)
Subreddit: CautiousBB	0.053 (0.117)
Subreddit: CBT	0.145 (0.108)
Subreddit: ChildrenofDeadParents	0.201 (0.128)
Subreddit: Codependency	0.052 (0.125)
Subreddit: CPTSD	0.112 (0.107)
Subreddit: cptsdcreatives	0.475 (0.295)
Subreddit: CPTSDNextSteps	0.185 (0.191)
Subreddit: datingoverfifty	0.265* (0.144)
Subreddit: datingoverforty	0.206* (0.111)
Subreddit: dbtselfhelp	0.134 (0.173)
Subreddit: death	**0.345** (0.162)**
Subreddit: DecidingToBeBetter	0.219* (0.114)
Subreddit: dementia	0.055 (0.136)
Subreddit: depression	0.157 (0.116)
Subreddit: domesticviolence	0.050 (0.128)
Subreddit: emotionalabuse	−0.191 (0.191)
Subreddit: emotionalneglect	0.184* (0.110)
Subreddit: ExNoContact	0.026 (0.123)
Subreddit: FriendsOver40	0.295 (0.221)
Subreddit: getdisciplined	**0.307** (0.124)**
Subreddit: grief	−0.088 (0.173)
Subreddit: GriefSupport	0.095 (0.108)
Subreddit: Grieving	0.273 (0.295)
Subreddit: happycryingdads	0.475 (0.295)
Subreddit: heartbreak	0.473 (0.295)
Subreddit: InternalFamilySystems	−0.038 (0.115)
Subreddit: IWantToLearn	**0.314*** (0.119)**
Subreddit: lastimages	−0.155 (0.131)
Subreddit: LifeAfterNarcissism	0.113 (0.134)
Subreddit: lonely	−0.038 (0.118)
Subreddit: MadeMeCry	0.475 (0.295)
Subreddit: marriageadvice	0.023 (0.221)
Subreddit: mentalhealth	0.068 (0.105)
Subreddit: MentalHealthUK	0.125 (0.107)
Subreddit: Miscarriage	0.076 (0.113)
Subreddit: MomForAMinute	0.066 (0.110)
Subreddit: NarcAbuseAndDivorce	0.098 (0.221)
Subreddit: narcissism	0.009 (0.191)
Subreddit: NarcissisticAbuse	0.089 (0.119)
Subreddit: NarcissisticSpouses	−0.177 (0.191)
Subreddit: OCD	0.167 (0.110)
Subreddit: offmychest	0.068 (0.106)
Subreddit: OldManDog	0.173 (0.162)
Subreddit: Petloss	0.116 (0.148)
Subreddit: PrayerRequests	0.275 (0.221)
Subreddit: PregnancyAfterLoss	0.121 (0.108)
Subreddit: productivity	0.258* (0.154)
Subreddit: psychotherapy	**0.234** (0.105)**
Subreddit: sad	−0.192 (0.295)
Subreddit: selfharm	0.183 (0.148)
Subreddit: selfimprovement	0.090 (0.121)
Subreddit: seniorkitties	0.223* (0.121)
Subreddit: SingleParents	0.103 (0.107)
Subreddit: socialanxiety	**0.368** (0.162)**
Subreddit: socialskills	**0.284** (0.113)**
Subreddit: SuicideBereavement	0.041 (0.125)
Subreddit: SuicideWatch	0.016 (0.115)
Subreddit: TalkTherapy	0.191* (0.105)
Subreddit: therapy	0.118 (0.105)
Subreddit: traumatoolbox	0.137 (0.173)
Subreddit: ttcafterloss	0.096 (0.119)
Subreddit: Vent	0.069 (0.173)
Subreddit: widowers	0.275 (0.173)
Constant	**0.527*** (0.104)**
Observations	20,029
R^2	0.077
Adjusted R^2	0.072
Residual Std. Error	0.276 (df = 19939)
F Statistic	18.560*** (df = 89; 19939)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 10: Regression results on predicting the percentage of alignment with title (categorical variable), subreddit (categorical variable), and whether title is visible or not. Model coefficients (β) are shown for each factor and standard errors are shown in parentheses. Bolded rows show factors that are significant at at least p<0.05.

Appraisal/Span		Subreddits
Pleasantness	Most misaligned	selfimprovement, getdisciplined, productivity, socialskills, askfuneraldirectors, CBT, Advice, PrayerRequests, IWantToLearn, AskPsychiatry
	Least misaligned	MadeMeCry, COVIDgrief, lastimages, Petloss, Miscarriage, GriefSupport, Grieving, widowers, happycryingdads, grief
Anticipated Effort	Most misaligned	gaslighting, sad, happycryingdads, marriageadvice, narcissism, IWantToLearn, askatherapist, askfuneraldirectors, Advice, AskPsychiatry
	Least misaligned	CPTSDNextSteps, ttcafterloss, FriendsOver40, widowers, PregnancyAfterLoss, BPDlite, Miscarriage, cptsdcreatives, CPTSD, MadeMeCry
Situational Control	Most misaligned	dbtselfhelp, MomForAMinute, NarcAbuseAndDivorce, psychotherapy, IWantToLearn, BodyAcceptance, marriageadvice, Advice, askfuneraldirectors, PrayerRequests
	Least misaligned	BPDlite, happycryingdads, CPTSD, widowers, Anxiety, BPD, ChildrenofDeadParents, depression, COVIDgrief, SuicideBereavement
Self-other Agency	Most misaligned	OCD, AskPsychiatry, askfuneraldirectors, MentalHealthUK, dbtselfhelp, PrayerRequests, IWantToLearn, getdisciplined, CautiousBB, productivity
	Least misaligned	NarcissisticAbuse, gaslighting, NarcissisticSpouses, emotionalabuse, BPDlovedones, abusiverelationships, marriageadvice, BreakUp, EmotionalAbuseSupport, ExNoContact
Certainty	Most misaligned	DecidingToBeBetter, selfimprovement, CancerCaregivers, happycryingdads, Anger, dbtselfhelp, getdisciplined, MentalHealthUK, productivity, IWantToLearn
	Least misaligned	death, OldManDog, lastimages, Petloss, heartbreak, grief, ChildrenofDeadParents, widowers, BreakUps, gaslighting
Objective Experience	Most misaligned	CBT, Anger, abusiverelationships, marriageadvice, domesticviolence, EmotionalAbuseSupport, Advice, emotionalabuse, gaslighting, PrayerRequests
	Least misaligned	CautiousBB, ttcafterloss, PregnancyAfterLoss, Miscarriage, COVIDgrief, ChildrenofDeadParents, lastimages, cancer, GriefSupport, Grieving
Advice	Most misaligned	CPTSDNextSteps, BorderlinePDisorder, BPD4BPD, widowers, NarcissisticAbuse, BPD, emotionalneglect, cptsdcreatives, BPDlite, CPTSD
	Least misaligned	IWantToLearn, Advice, MentalHealthUK, PrayerRequests, CancerCaregivers, AskPsychiatry, CBT, MomForAMinute, selfimprovement, MMFB

Table 11: Ten most misaligned and ten least misaligned subreddits for each appraisal/span.

Desciption	Example	Aligned? Model	Aligned? Label
Overgeneralization	**Target:** IDK, I guess despair is just all over in me. **Observer:** When things go so badly that you want to see it become even worse.	Yes	No
Pattern-overcatching (First, Second, ...)	**Target:** I don't really know what to do. **Observer:** Second, any thoughts you're having are nothing to feel guilty about or be as homes of.	Yes	No
Wrong object ("he" and "I")	**Target:** Am I in the wrong? Do I need therapy to help me get over his past hurtful behavior... **Observer:** You aren't obligated to feel or act a certain way to make him feel connected to you or lessen any guilt he may feel.	Yes	No
Implicit reference	**Target:** I feel like I'm alone all the time so I might as well just be alone **Observer:** I feel like you are so occupied with what you don't have, you're thinking about what you donn't have.	No	Yes
Explicit reference	**Target:** But why does it still hurt.. **Observer:** It takes a long time sometimes to get over someone.	No	Yes
Valid alignment in experience	**Target:** She broke up with me, she destroyed my heart and she's the one finding someone new to be happy with while i haven't had even a crush on anyone in all this time. **Observer:** i also felt like i was loosing in the breakup because i could not move on as fast	No	Yes

Table 12: Examples of errors that our best model makes when predicting alignment.