Curating a Large-Scale Motivational Interviewing Dataset using Peer Support Forums

Anuradha Welivita and **Pearl Pu**
School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne
Switzerland
{kalpani.welivita,pearl.pu}@epfl.ch

Abstract

A significant limitation in developing therapeutic chatbots to support people going through psychological distress is the lack of high-quality, large-scale datasets capturing conversations between clients and trained counselors. As a remedy, researchers have focused their attention on scraping conversational data from peer support platforms such as Reddit. But the extent to which the responses from peers align with responses from trained counselors is understudied. We address this gap by analyzing the differences between responses from counselors and peers by getting trained counselors to annotate ≈17K such responses using Motivational Interviewing Treatment Integrity (MITI) code, a well-established behavioral coding system that differentiates between favorable and unfavorable responses. We developed an annotation pipeline with several stages of quality control. Due to its design, this method was able to achieve 97% of coverage, meaning that out of the 17.3K responses we successfully labeled 16.8K with a moderate agreement. We use this data to conclude the extent to which conversational data from peer support platforms align with real therapeutic conversations and discuss in what ways they can be exploited to train therapeutic chatbots.

1 Introduction

Demands of the modern world are increasingly responsible for bringing adverse impacts on our mental health. World Health organization estimates psychological distress affects 29% of people in their lifetime (Steel et al., 2014). Shortage of mental health workers and the stigma associated with mental health further demotivates people in actively seeking out help. Thus, provision of mental health support through the use of AI-driven conversational agents to complement traditional therapy has become an interesting area of research (Fitzpatrick et al., 2017; Inkster et al., 2018; Mousavi et al., 2021). But one challenge associated with developing such agents is the lack of large-scale psychotherapeutic conversations. They are either limited or are not available publicly due to ethical reasons.

Nowadays, with the expansion of social media, it could be observed that people use social media platforms such as Reddit to vent their distress and peers are seen to actively respond to such posts. These conversations are available in abundance and are publicly accessible through web scraping APIs. Thus, conversations scraped from such platforms are seen as an alternative to overcome the above challenge (Welivita and Pu, 2022). Prior work has found that responses from peers contain higher empathic concern for posts for seeking help as many peers share similar distressful experiences (Hodges et al., 2010). But the extent to which responses from peers align with responses from trained counselors remain a major limitation. Knowing these differences can shed light on to what extent such conversational data could be used in training therapeutic chatbots and what pre-processing or rephrasing steps that one should take if they are being used for such purposes.

Though studies have been conducted independently to assess the competency of counselors and peers offering support (Pérez-Rosas et al., 2016; Klonek et al., 2015; Gaume et al., 2009; Sharma et al., 2020a; De Choudhury and De, 2014), studies that comparatively analyse the differences between them are limited. One such study was conducted by Lahnala et al. (2021), where they show that a classifier can distinguish between responses provided to help-seeking posts regarding mental health by professionals and peers. Mousavi et al. (2021) conducted a similar analysis between responses collected from psychotherapists and non-expert dialogue writers and noted linguistic variability in the two types of responses. However, all these analyses are limited to the lexical level.

To address the above gaps, we comparatively an-

Proceedings of the 29th International Conference on Computational Linguistics, pages 3315–3330
October 12–17, 2022.

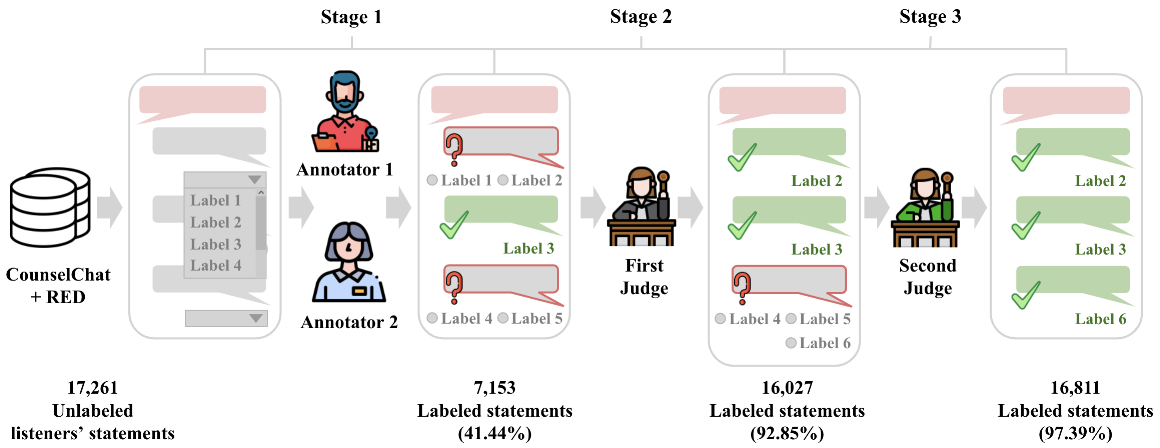


Figure 1: The annotation process to label the listeners' statements in the CounselChat and RED datasets with labels adapted from MITI. The process was conducted in three main stages.

alyze responses from professional counselors and peers by annotating these responses using labels adapted from a well-established behavioral coding scheme named Motivational Interviewing Treatment Integrity (MITI) code (Moyers et al., 2014). The MITI code is used in psychological literature to evaluate how well a mental health practitioner responds to those seeking help with their mental health related issues. Specific response types from the MITI code have shown to increase the likelihood of positive health outcomes (Pérez-Rosas et al., 2018; Gaume et al., 2009). For this purpose, we make use of post-response pairs scraped from the CounselChat website (counselchat.com) that contains high-quality therapist responses to emotional distress related questions from individuals and dialogues curated from several mental-health related subreddits in Reddit, in which peers engage in actively to respond to those seeking help.

Annotating dialogue responses with labels from the MITI coding system is known to be very time consuming and expensive as it requires expert annotators trained in the practice of psychology and careful attention to the labelling task (Pérez-Rosas et al., 2016). This human labour is difficult to find. But the availability of crowdsourcing platforms such as UpWork [1] and Fiverr [2] have made it more accessible to find human labour that satisfy our requirements. Thus, we were able to recruit professionally trained mental health practitioners through UpWork to annotate dialogue responses with labels adapted from the MITI code. Our annotation pipeline consisted of three stages as depicted in Figure 1. Two workers were involved in the first stage and high-quality workers who scored high observed agreements with a peer in the first stage acted as judges to resolve conflicting labels in the second and third stages contributing to increased observed agreement and inter-rater reliability.

Using these annotations, we conducted a comparative analysis between responses from peers and counselors to identify to what extent they align with each other. Based on these findings, we recommend ways of boosting peers' responses to match as close as possible to counselors' responses. The recommendations made in this paper can contribute to improve the perceived therapeutic effectiveness of a chatbot trained on data from peer support forums.

Our contributions are three fold. 1) We develop an MI dataset having client-counselor and peer-peer dialogues, in which ≈17K listeners' utterances are annotated with labels adapted from the MITI code.[3] 2) We discuss the details of the annotation process followed in increasing the agreement between the workers when annotating with MITI codes. 3) Based on these annotations, we conduct a comparative analysis between counselors' and peers' responses and draw useful conclusions on to what extent responses from peers align with responses from trained counselors and recommend ways of boosting peers' responses such that it can increase the perceived effectiveness of therapeutic chatbots trained on such data.

2 Related Work

Motivational Interviewing Treatment Integrity (MITI) is a behavioral coding system that measures

[1] http://upwork.com
[2] https://www.fiverr.com
[3] The dataset can be downloaded at https://github.com/anuradha1992/Motivational-Interviewing-Dataset.

how well a mental health practitioner uses motivational interviewing in therapy (Moyers et al., 2003). It exclusively focuses on the verbal behaviour of a counselor and is used to increase clinical skill in the practice of motivational interviewing (MI). This coding system has been used extensively to to improve the understanding of the counseling practice alone (Can et al., 2012; Pérez-Rosas et al., 2018, 2019). Pérez-Rosas et al. (2016) developed an MI dataset consisting of ≈22K counselors' responses during Motivational Interviewing encounters annotated with 10 behavioral codes from the MITI. Althoff et al. (2016) conducted a quantitative study on counseling conversations to measure how various linguistic aspects of conversations are correlated with conversation outcomes. However, the datasets used in such analyses are not publicly available due to ethical reasons. Thus, it is difficult to use such resources in training therapeutic chatbots even though real counselor responses are the ideal candidates for this purpose.

There are a number of research efforts taken to develop therapeutic chatbots (Fitzpatrick et al., 2017; Inkster et al., 2018; Welch et al., 2020; Mousavi et al., 2021). Among them, recent work focuses on using dialogue data from peer-support forums (Sharma et al., 2020b; Welivita and Pu, 2022). For example, Welivita and Pu (2022) developed a knowledge-graph containing distress consoling responses for specific types of stressors using peer-support responses from Reddit and utilized it in single-turn dialogues. Some studies specifically focus on attributes such as perceived empathy and information richness in mental health related discourse in peer support forums that suggests they are good candidates for training such chatbots (Nambisan, 2011; De Choudhury and De, 2014; Sharma et al., 2020a,b). But these studies lack comparisons with responses generated by professional counselors. In our work, we mainly attempts to address this limitation.

3 Methodology

In this section, we describe the methodology including the labels chosen to annotate the listeners' statements, the datasets used, and different stages of the annotation process.

3.1 Labels Adapted from MITI

The labels we used for annotation were adapted from MITI code 2.0 (Moyers et al., 2003) and 4.2.1 (Moyers et al., 2014). Table 1 shows the MITI labels that were used for annotation with descriptions and examples. Altogether, there are 15 labels. They also include labels *Self-Disclose* and *Other*, which are not included in the MITI code. We included the label *Self-Disclose* because in conversations involving peer support, *Self-Disclosure* is an important aspect that enables the sharing of lived experience and is seen to occur quite frequently in the majority of such conversations (Truong et al., 2019). The above labels were used to annotate each sentence in the listeners' utterances.

3.2 Datasets

We used two datasets containing distress consoling dialogues for annotation: 1) *CounselChat* that contains responses from professional counselors; and 2) *RED* that contains responses from peers.

CounselChat Dataset: The CounselChat dataset consists of high-quality therapist responses to emotional distress related questions from individuals. This data is scraped from the CounselChat website (counselchat.com), which is an online platform that helps counselors to make meaningful contact with potential clients. On the website, professional counselors respond to questions posed by users suffering from mental health issues and emotional distress. The dataset consists of 2,129 post-response pairs that span across 31 distress related topics, the most frequent topics being *depression*, *relationships*, and *intimacy*. This dataset is publicly available. [4] Out of this data, we randomly selected 1,000 post-response pairs to be annotated with labels derived from the MITI code.

Reddit Emotional Distress (RED) Dataset: To obtain a dialogue dataset containing utterances from peer-supporters as response for posts containing emotional distress, we scraped a new dataset from Reddit, containing dialogues that discuss real-world distressful situations. Reddit was chosen since it is publicly available and peers actively engage in Reddit to support others going through distressful situations in life. We used the Pushshift API (Baumgartner et al., 2020) to collect and process dialogue threads from a carefully selected set of 8 subreddits: *mentalhealthsupport*; *offmychest*; *sad*; *suicidewatch*; *anxietyhelp*; *depression*; *depressed*; and *depression_help*, which are popular among Reddit users to vent their distress. We ex-

[4] `https://github.com/nbertagnolli/counsel-chat`

MITI label	Description	Examples
1. Closed Question	Questions that can be answered with an yes/no response or a very restricted range of answers.	*Do you think this is an advantage?*
2. Open Question	Questions that allow a wide range of possible answers.	*What is your take on that?*
3. Simple Reflection	Repetition, rephrasing, or paraphrasing of speaker's previous statement.	*It sounds like you're feeling worried.*
4. Complex Reflection	Repeating or rephrasing the previous statement of the speaker but adding substantial meaning/emphasis to it.	**Speaker:** *Mostly, I would change for future generations.* **Listener:** *It sounds like you have a strong feeling of responsibility.*
5. Give Information	Educating, providing feedback, or giving an opinion without advising.	*Logging your cravings is important as cravings often lead to relapses.*
MI Adherent Behaviour Codes:		
6. Advise with Permission	Advising when the speaker asks directly for advice. Indirect forms of permission can also occur, such as when the listener says to disregard the advice as appropriate.	*If you agree with it, we could try to brainstorm some ideas that might help.*
7. Affirm	Encouraging the speaker by saying something positive or complimentary.	*You should be proud of yourself for your past's efforts.*
8. Emphasize Autonomy	Emphasizing the speaker's control, freedom of choice, autonomy, and ability to decide.	*It is really up to you to decide.*
9. Support	Statements of compassion or sympathy.	*I know it's really hard to stop drinking.*
MI Non-Adherent Behaviour Codes:		
10. Advise without Permission	Making suggestions, offering solutions or possible actions without first obtaining permission from the speaker.	*You should simply scribble a note that reminds you to take a break.*
11. Confront	Directly and unambiguously disagreeing, arguing, blaming, criticizing, or questioning the speaker's honesty.	*Yes, you are an alcoholic. You might not think so, but you are.*
12. Direct	Giving orders, commands, or imperatives.	*Don't do that!*
13. Warn	A statement or event that warns of something or that serves as a cautionary example.	*Be careful, DO NOT stop taking meds without discussing with your doctor.*
Other:		
14. Self-Disclose	The listener discloses his/her personal information or experiences.	*I used to be similar where I get obsessed about how people look.*
15. Other	Statements that are not classified under the above codes	*Good morning, Hi there.*

Table 1: The set of labels adapted from the MITI code, which were used to annotate listeners' responses.

tracted the dialogue turns out of these threads and subjected these dialogues to a rigorous data cleaning pipeline, which included removal of profanity from listener responses. By this, we were able to curate $\approx 1.2M$ dyadic conversations having on average 2.66 turns per dialogue. Out of them, 1K dialogues were randomly selected for annotation.

3.3 Annotation Experiment

We used UpWork, a leading crowdsourcing platform to recruit trained counselors to annotate dialogue responses from CounselChat and RED datasets. Altogether 12 workers were recruited to annotate 2K dialogues, 1K from CounselChat and 1K from RED. They contained 17,261 individual sentences in the listener utterances in total.

The task was carried out in three stages. First, the workers were asked to annotate each sentence contained in the listener utterances of the dialogues from CounselChat and RED datsets with one of fifteen MITI labels. We bundled ten dialogues into one batch (a batch contained five CounselChat and five RED dialogues interchangeably) and assigned two workers per batch. At the beginning, a tutorial about the labels accompanied by a practice task was offered to self-validate the workers' answers. As the task was ongoing, we computed the observed agreement of each worker with peers and offered more batches for the workers whose observed agreement was better than the others.

At the end of stage 1, we noticed that the two workers assigned for each batch did not agree on a common label for more than half of the listeners' sentences in the two datasets. Manual inspection of their answers revealed majority of the disagreements occurred because there are highly confusing pairs of labels that need more careful attention to differentiate (e.g. *Complex Reflection* and *Give Information* can be easily confused). Hence, we launched a second stage of the experiment by asking four workers who scored the highest observed agreement with a peer in the first stage to act as judges to resolve these conflicting labels. A judge was presented with the two labels the workers specified in the first stage along with the listener's sentence and the dialogue context and was asked to select either one of the two labels if one of them agreed with the listener's sentence. Only if none of the labels agreed with the listener's sentence, he was instructed to select a label from the rest.

At the end of stage 2, most of the conflicting labels were resolved by the judge's annotations. But there was a small percentage of listeners' sentences for which a label was still not agreed upon. We noticed for 68.15% of such unresolved sentences, at least one annotation was given by a relatively poor performing worker whose observed agreement score with a peer was below average as measured in the first stage. We decided such labels are not worth considering since they cloud the decision process and chose to launch a third stage of the experiment by removing one annotation given by the poorest performing worker for each such unresolved sentence. Similar to stage 2, we recruited the same judges and presented them with the two remaining labels to be resolved. This entire annotation pipeline is illustrated in Figure 1.

4 Results

Table 2 shows the statistics of the results from each stage of the experiment. At the end of stage 1, out of 17,261 listeners' sentences, 7,152 received a label as agreed by the two annotators. Altogether, by end of stage 1, we could yield an observed agreement percentage of 41.43% and an inter-rater agreement (Fleiss' kappa) score of 0.3391 indicating fair agreement. At the end of stage 2, another 8,875 labels were resolved, yielding an observed agreement of 87.79%. The updated inter-rater agreement (Fleiss' kappa) after this stage was 0.5292, which is a significant increase compared to the previous stage. After the end of completion of stage 1 and stage 2 of the annotation process, from among the total of 17,261 listeners' sentences in CounselChat and RED datasets, 16,027 of them were able to receive a label as agreed by at least two workers. This is 92.85% of the entire data.

From the remaining 1,234 sentences for which a label was not agreed upon, 841 (68.15%) sentences were annotated by at least one poor worker whose observed agreement with a peer was below average. At the end of stage 3 of the experiment, which was conducted by removing such annotations given by the poor workers, a second judge was able to agree with one of the two remaining labels for 784 sentences, yielding an observed agreement of 93.22%. The updated inter-rater agreement (Fleiss' kappa) after the third stage was 0.5453 (moderate agreement), showing a slight increase compared to the score in the previous stage. The lower kappa scores are potentially due to the inherent difficulty of distinguishing some of the MI labels, which we elaborate below. A similar annotation experiment conducted by Perez-Rosas et al. (2016) report similar kappa scores ranging from 0.31 to 0.64 on different MI labels.

The confusion matrices computed at each stage of the experiment are denoted in the appendices. We could observe that the label pair *Complex Reflection - Give Information* had the highest percentage of disagreement between the two workers in both CounselChat and RED datasets. In addition, the label pairs *Advise with Permission - Advise without Permission* and *Give Information - Advise without Permission* were highly confusing to differentiate in the CounselChat dataset. Whereas, in RED, the label pairs *Affirm - Support* and *Advise without Permission - Direct* were difficult to be differentiated. These observations were quite intuitive since these pairs of labels either contained similar semantic content (e.g. *Complex Reflection - Give Information*, *Advise with Permission - Advise without Permission*, *Give Information - Advise without Permission*, *Advise without Permission - Direct*) or used similar language constructs (e.f. *Affirm - Support*, *Advise without Permission - Direct*).

Final aggregated statistics of the three stages of the annotation process is shown in Table 3. It could be observed how the labels grew to cover a larger portion of the listeners' sentences as the annotation process advanced through the stages. Finally, close to 97% of the listeners' sentences (16,812 in total) were annotated with the MITI labels.

5 Analysis of the MI Dataset

In Figure 2, we show the distribution of labels adapted from the MITI code in CounselChat and RED datasets, separately. Table 4 shows the specific number of each label in CounselChat and RED datasets and the increase/decrease in each label in the two datasets compared to each other. Based on these statistics, we discuss seven major differences

Description	CC	RED	CC + RED
Stage 1:			
Total number of listeners' sentences annotated	9,893	7,368	17,261
Sentences for which a label was agreed upon by both annotators	4,067	3,085	7,152
Observed agreement between the two annotators	41.11%	41.87%	41.43%
Inter-rater agreement (Fleiss' kappa)	0.3059	0.3577	0.3391
Stage 2:			
The number sentences, for which, the label had to be resolved	5,826	4,283	10,109
The number of times the judge agreed with one of the given labels	5,111	3,764	8,875
Observed agreement between the judge and one of the two annotators	87.73%	87.88%	87.79%
Updated inter-rater agreement (Fleiss' kappa)	0.5029	0.5440	0.5292
Stage 3:			
The number sentences, for which, the label had to be resolved	479	362	841
The number of times the judge agreed with one of the given labels	450	334	784
Observed agreement b/w the judge and one of the remaining annotators	93.95%	92.27%	93.22%
Updated inter-rater agreement (Fleiss' kappa)	0.5193	0.5601	0.5453

Table 2: Statistics of the three stages of the annotation experiment. The CounselChat dataset is abbreviated as CC.

Description	CC	RED	CC + RED
# listener sentences	9,893	7,368	17,261
# labels agreed in stage 1	3,085 (41.11%)	4,067 (41.87%)	7,152 (41.43%)
# labels agreed in stage 2	9,178 (92.96%)	6,849 (92.77%)	16,027 (92.85%)
# labels agreed in stage 3	9,628 (97.49%)	7,183 (97.32%)	16,811 (97.39%)

Table 3: Final aggregated statistics of the three stages of the annotation process.

observed between responses from counselors and peers and state our recommendations when using this data to train therapeutic conversational agents.

Questions: There is ≈23% and ≈26% increase in closed and open questions, respectively, in counselor responses from CC compared to peer-support responses from RED. Questioning plays a central role in therapeutic interactions as it builds up mutual dialogue between client and therapist (Poskiparta et al., 2000). But Hill et al. (1983) observed with time-limited counseling, fact finding through closed questions is rated lower in helpfulness. It can result in the speaker saying less and less and the listener feeling pressured to ask more questions to keep the interaction going. However, in both CounselChat and RED datasets, the number of open questions is nearly half of the number of closed questions. Hence, mechanisms should be devised to increase the percentage of open questions to balance the number of closed questions. This combination would be more effective than a disproportionate reliance on closed questions.

Reflections: The number of reflections is positively associated with the perceived empathy (Klonek et al., 2015). It is also a competence indicator in assessing MI competency (Moyers et al., 2003). Non-surprisingly, both simple and complex reflections are observed to be higher (≈20% and ≈30% increase in simple and complex reflections, respectively) in counselors' responses compared to peers'. Thus, it would be beneficial to boost the percentage of reflections among peer support dialogues when using them to train therapeutic agents.

Scholars emphasize that listeners should formulate more reflections than closed questions (Klonek et al., 2015). As we observed, some closed questions such as *"Are you eating because you are bored?"* are identical to reflections, differing only in the voice intonation at the end. They could be easily reformulated into reflections such as *"It seems that you are eating because you are bored"*.

Giving information: In counselor responses, there is a 200.33% increase of *Give Information* type of sentences compared to responses from peers. It is quite unsurprising since counselors are relatively knowledgeable about the subject being discussed and hence are in a position to provide information that can help the speaker. Informed by this observation, steps should be taken to boost the amount of information in peer-support responses.

MI Adherent Behavior Codes: Supporting the client with statements of compassion and sympathy are surprisingly higher among peers (≈95% increase) compared to counselors. Affirming the speaker by saying something positive or complimentary is also seen to be comparatively higher in RED (≈21% increase). These are very good indicators that show peer-support responses if uti-

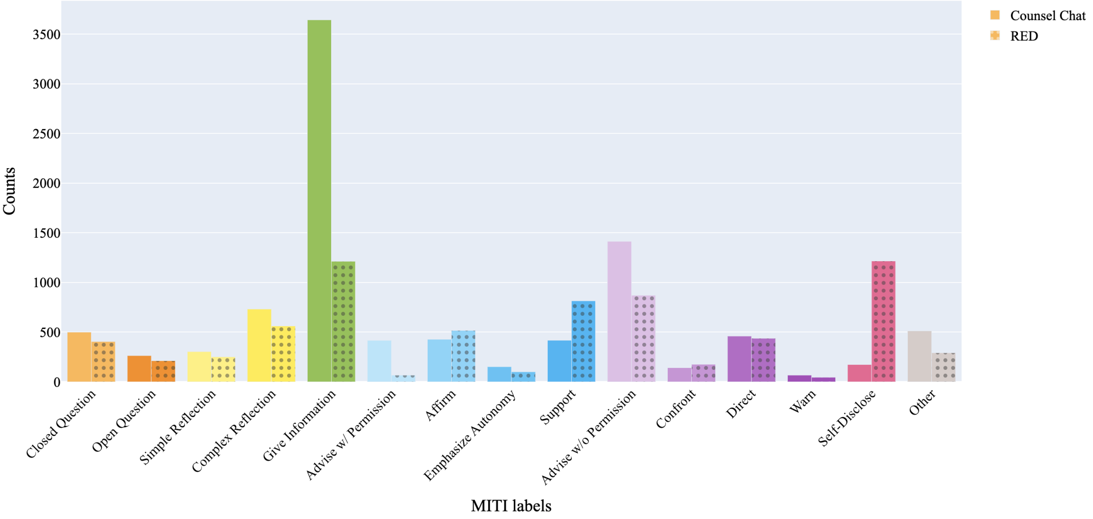


Figure 2: Distribution of MITI labels in CounselChat and RED datasets.

Label	**No. of labels in CounselChat**	**No. of labels in RED**	**Increase in CC compared to RED**	**Increase in RED compared to CC**
Closed Question	500	405	23.46% ↑	-19.00% ↓
Open Question	264	212	24.53% ↑	-19.70% ↓
Simple Reflection	304	252	20.63% ↑	-17.11% ↓
Complex Reflection	732	562	30.25% ↑	-23.22% ↓
Give Information	3,643	1213	200.33% ↑	-66.70% ↓
MI Adherent Behavior Codes:				
Advise with Permission	417	67	522.39% ↑	-83.93% ↓
Affirm	428	517	-17.21% ↓	20.79% ↑
Emphasize Autonomy	152	101	50.50% ↑	-33.55% ↓
Support	418	815	-48.71% ↓	94.98% ↑
MI Non-Adherent Behavior Codes:				
Advise without Permission	1,414	871	62.34% ↑	-38.40% ↓
Confront	142	176	-19.32% ↓	23.94% ↑
Direct	460	438	5.02% ↑	-4.78% ↓
Warn	67	46	45.65% ↑	-31.34% ↓
Other:				
Self-Disclose	174	1216	-85.69% ↓	598.85% ↑
Other	513	292	75.68% ↑	-43.08% ↓

Table 4: Statistics of MITI labels in CounselChat and RED datasets and the increase/decrease in each label in the two datasets compared to each other. The increases/decreases that are favourable for the interaction are indicated in green while those that are unfavourable are indicated in red. The increases/decreases in *Self-Disclose* and *Other* are not assigned a color as their role in therapeutic interventions are quite blurry and subjected to debate.

lized in training therapeutic agents will reflect more compassion, sympathy, positivity, and compliments towards the user in distress. On the other hand, emphasizing the speaker's control and autonomy is observed to be higher in counselors' responses (≈50% increase) compared to responses from peers.

Advising with and without permission: Giving advices is generally seen to be higher in counselor responses. There is ≈522% increase in advising after asking for permission and ≈62% increase in advising without asking for permission among counselors' responses compared to those from the peers. Advising without asking for permission takes a portion of 77.22% of the total number of advices given in counselor responses. Thus, counselors, though professionally trained, tend to make the mistake of advising without prior asking for the speaker's permission. This percentage is higher in peer-support responses in which advising without permission takes a portion of 92.86% of the total number of advices given by the peers. Thus, in both datasets, steps should be taken to reformulate advices in a way that the agent asks for the speakers' permission before giving advice.

MI Non-Adherent Behavior Codes: Confronting the client by directly disagreeing, arguing, or criticizing is higher in peers' responses (≈24% increase) compared to those of the counselors'.

Such interactions reflect uneven power sharing, accompanied by disapproval and negativity (Moyers et al., 2003). Directing the speaker by giving orders and also warnings are quite surprisingly seen to be slightly higher in the responses given by the counselors compared to the responses of the peers (≈5% and ≈46% increase for *Direct* and *Warn*, respectively). These are non-favourable response types that negatively affect the therapeutic interaction between the speaker and the listener and thus should be detected and eliminated as a preliminary step before using such responses to train chatbots.

Self-Disclosure: The role of self-disclosure in therapeutic interventions is quite blurry. For example, psychoanalysts believe that self-disclosure is counterproductive as it distorts client's transference. Conversely, Cognitive Behavioural therapists believe that self-disclosure can be a useful tool in therapy as it models and reinforces new perspectives for the client. Digging deep, there are two broad types of self-disclosure used by counsellors: 1) *intra-session disclosure*, where the counselor discloses a feeling about the client that is relevant to the therapeutic process; and 2) *extra-session disclosure*, where the counselor reveals information about themselves that occurs outside the session. In most cases, *intra-session disclosure* is the most useful type of self-disclosure (Levitt et al., 2016).

As we manually inspected the statements labeled as *Self-disclosure* in CounselChat and RED datasets, it was found out that *Intra-session disclosure* is seen higher in CC compared to RED, whereas *Extra-session disclosure* is seen higher in RED compared to CC. Table 5 provides some examples of such statements. This suggests that counselors are more careful when disclosing information about themselves and when they do they make sure that the information they disclose is relevant to the therapeutic process. Extra-session disclosure too has its place in therapeutic interactions specially contributing to building rapport between the client and the therapist. However, as suggested by R. Schwartz (2021), this type of disclosure must be used wisely with caution since it can as well be counterproductive distorting client's transference.

Examples of *intra-session disclosure* in CounselChat:
- Personally, I can tell you that I would want my clients to tell me about anxiety they feel 100% of the time. *- I have had clients asking the same question and there is often an underlying fear that they "can't be helped" or they will "be too much for their therapist."*
Examples of *extra-session disclosure* in RED:
- You remind me a lot of my best friend that I had when I was young. Being her friend was exhausting. *- I too suffer from psychosis from my schizo-affective disorder, yelled at my former best friend for gangstalking me, called her all kinds of horrible names.*

Table 5: Examples of different types of self-disclosure observed in CounselChat and RED datasets.

6 Discussion and Conclusion

This paper discussed the curation process of a large-scale distress consoling dialogue dataset containing utterances from trained counselors and peers. A carefully designed annotation process was followed to annotate each response statement with labels adapted from the MITI code. We saw the effectiveness of our annotation process as it contributed in increasing the observed agreement and inter-rater reliability as the process advanced through different stages. Based on the comparative analysis between responses from counselors and peers, we reported seven major differences between them, highlighting the strengths and limitations of using abundantly available peer-support dialogues for purposes such as training therapeutic chatbots. In summary, peers' responses tend to be more supportive, compassionate, and encouraging than counselors' as observed by the increased percentage of *Support* and *Affirm* labels. But important therapeutic techniques such as asking more open questions than closed ones, reflections, giving information and advices with permission, and emphasizing speaker's autonomy require further boosting. MI non-adherent behaviors such as confronting is also seen higher among peers and thus should be eliminated. Careful attention should also be paid to self-disclosure among peers as the majority of such statements are of the type *extra-session disclosure*, which is less useful for the therapeutic process.

Curating this dataset is the first step in our general goal of boosting the therapeutic competency in peer-support responses. Using this dataset we plan to train an MITI classifier to automatically identify favourable and unfavourable response types present in peers' responses. Being able to detect MI non-adherent behaviors such as confronting will enable us to directly eliminate such responses from the data. Next, we intend to develop an MITI rephraser that can convert certain types of responses such as closed questions and advices without permission into more favourable reflections and advices with permission, respectively. We plan on investigating simple linguistic rule based approaches as well as unsupervised text style transfer methods that can be trained on unparalleled corpora (Malmi et al., 2020; Jin et al., 2022) for this purpose. We believe this will largely boost the therapeutic competency in peer-support responses and will increase the therapeutic effectiveness in chatbots trained on them.

7 Ethics Statement

Data curation: Analysis of posts of a website such as Reddit is likely considered "fair play" as individuals are anonymous and users are aware that their responses remain archived on the site unless they explicitly delete them. The Reddit privacy policy also states it allows third parties to access public Reddit content through the Reddit API and other similar technologies. [5] Reddit's data is already widely available in larger dumps such as Pushshift (Baumgartner et al., 2020). We collected only publicly available data in Reddit and the curation process did not involve any intervention or interaction with the Reddit users. The CounselChat dataset is also available publicly. But Fiesler and Proferes (2018) in a study on user perceptions on social media research ethics empahsizes some potential harms that can be caused due to social computing research because internet users rarely read or could fully understand website terms and conditions. Since this dataset in particular contains sensitive information, we adhere to the guidelines suggested by Benton et al. (2017) for working with social media data in health research, and share only anonymized and paraphrased excerpts from the dataset so that it is not possible to recover usernames through a web search with the verbatim post text. In addition, references to usernames as well as URLs are removed from dialogue content for de-identification.

Human annotation: Considering the qualifications of the workers we recruited, who were all trained in the practice of counseling, we were determined to pay them a wage considerably above the US minimum wage of $7.12 per hour. We paid them $10 per batch of 10 dialogues in the first stage of the experiment, in which average task completion time took ≈30 minutes (excluding the time taken by workers who took an unusually long time to complete the task). A bonus of $5 was offered for each batch if a worker obtained an above average observed agreement with a peer. For the second and third stages of the annotation task, we offered the workers 5 per batch of 10 dialogues. The average completion time per batch was ≈15 minutes in these two stages. Since the dataset is in English, all the annotators recruited were either native speakers or had professional competency in the English language. The fact that the dataset is English-only potentially perpetuates an English bias in NLP systems.

Therapeutic chatbots: Finally, there can be certain ethical implications associated with the development of therapeutic chatbots as highlighted by several researchers (Lanteigne, 2019; Montemayor et al., 2021; Tatman, 2022). However, the idea of therapeutic chatbots is not a new concept. Chatbots such as SimSensei (DeVault et al., 2014), Dipsy (Xie, 2017), Emma (Ghandeharioun et al., 2019), Woebot (`woebothealth.com`), and Wysa (`www.wysa.io`) are some examples. As Czerwinski et al. (2021) state, *About 1 billion people globally are affected by mental disorders; a scalable solution such as an AI therapist could be a huge boon*. Thus, even though therapeutic chatbots may encompass certain ethical implications, based on previous studies we already can acknowledge that the use of chatbots has the potential to improve therapeutic services notably in relation to accessibility and anonymity.

We curated this dataset for the ultimate development of a chatbot that adheres to MI strategy when responding to emotional distress. With the significant performance achieved by recent pre-trained language models, going for a deep learning-based solution is one of the choices that can be taken when developing such an agent. But it should not be undermined that because of the unpredictability associated with generative models, they always carry a risk when delivering emotional support to those undergoing distress. Thus, caution should be taken to avoid the delivery of inappropriate responses. This may not be limited to avoiding profane or judgemental responses. As pointed out by R. Tatman (2022) a response such as *"You're not alone"* may be comforting to someone with depression, however, can bring detrimental effects to someone suffering from paranoia. Hence, caution should be taken when developing therapeutic chatbots based on this dataset. Real-world deployment of such agents may still encompass potential dangers and if deployed, should be done with human supervision.

[5] `www.redditinc.com/policies/privacy-policy-october-15-2020`

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102.

Doğan Can, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2012. A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Mary Czerwinski, Javier Hernandez, and Daniel McDuff. 2021. Building an ai that feels: Ai systems with emotional intelligence could learn faster and be more helpful. *IEEE Spectrum*, 58(5):32–38.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1).

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.

Jacques Gaume, Gerhard Gmel, Mohamed Faouzi, and Jean-Bernard Daeppen. 2009. Counselor skill influences outcomes of brief motivational interventions. *Journal of substance abuse treatment*, 37(2):151–159.

Asma Ghandeharioun, Daniel McDuff, Mary Czerwinski, and Kael Rowan. 2019. Emma: An emotion-aware wellbeing chatbot. In *International Conference on Affective Computing and Intelligent Interaction*.

Clara E Hill, Jean A Carter, and Mary K O'Farrell. 1983. A case study of the process and outcome of time-limited counseling. *Journal of Counseling Psychology*, 30(1):3.

Sara D Hodges, Kristi J Kiel, Adam DI Kramer, Darya Veach, and B Renee Villanueva. 2010. Giving birth to empathy: The effects of similar experience on empathic accuracy, empathic concern, and perceived empathy. *Personality and Social Psychology Bulletin*, 36(3):398–409.

Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. 2018. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Florian E Klonek, Vicenç Quera, and Simone Kauffeld. 2015. Coding interactions in motivational interviewing with computer-software: What are the advantages for process researchers? *Computers in Human Behavior*, 44:284–292.

Allison Lahnala, Yuntian Zhao, Charles Welch, Jonathan K. Kummerfeld, Lawrence C An, Kenneth Resnicow, Rada Mihalcea, and Verónica Pérez-Rosas. 2021. Exploring self-identified counseling expertise in online support forums. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4467–4480, Online. Association for Computational Lingfgfggftzr666757tl.uistics.

Camylle Lanteigne. 2019. Social robots and empathy: The harmful effects of always getting what we want.

Heidi M Levitt, Takuya Minami, Scott B Greenspan, Jae A Puckett, Jennifer R Henretty, Catherine M Reich, and Jeffery S Berman. 2016. How therapist self-disclosure relates to alliance and outcomes: A naturalistic study. *Counselling Psychology Quarterly*, 29(1):7–28.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8671–8680, Online. Association for Computational Linguistics.

Carlos Montemayor, Jodi Halpern, and Abrol Fairweather. 2021. In principle obstacles for empathic ai: why we can't replace human empathy in healthcare. *AI & society*, pages 1–7.

Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. 2021. Would you like to tell me more? generating a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9.

TB Moyers, JK Manuel, D Ernst, T Moyers, J Manuel, D Ernst, and C Fortini. 2014. Motivational interviewing treatment integrity coding manual 4.1 (miti 4.1). *Unpublished manual*.

Theresa B Moyers, Tim Martin, Jennifer K Manuel, William R Miller, and D Ernst. 2003. The motivational interviewing treatment integrity (miti) code: Version 2.0. *Retrieved from Verfübar unter: www. casaa. unm. edu [01.03. 2005]*.

Priya Nambisan. 2011. Information seeking and social support in online health communities: impact on patients' perceived empathy. *Journal of the American Medical Informatics Association*, 18(3):298–304.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51.

Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 2018. Analyzing the quality of counseling conversations: the tell-tale signs of high-quality counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935.

Marita Poskiparta, Tarja Kettunen, and Leena Liimatainen. 2000. Questioning and advising in health counselling: results from a study of finnish nurse counsellors. *Health Education Journal*, 59(1):69–89.

Robert Schwartz. 2021. The big reveal | ethical implications of therapist self-disclosure.

Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020a. Engagement patterns of peer-to-peer interactions on mental health platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 614–625.

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

Zachary Steel, Claire Marnane, Changiz Iranpour, Tien Chey, John W Jackson, Vikram Patel, and Derrick Silove. 2014. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International journal of epidemiology*, 43(2):476–493.

Rachael Tatman. 2022. [link].

C Truong, J Gallo, D Roter, and J Joo. 2019. The role of self-disclosure by peer mentors: Using personal narratives in depression care. *Patient education and counseling*, 102(7):1273–1279.

Charles Welch, Allison Lahnala, Veronica Perez-Rosas, Siqi Shen, Sarah Seraj, Larry An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2020. Expressive interviewing: A conversational system for coping with covid-19. *arXiv preprint arXiv:2007.03819*.

Anuradha Welivita and Pearl Pu. 2022. Heal: A knowledge graph for distress management conversations.

Xing Xie. 2017. Dipsy: A digital psychologist.

A Annotation Experiment

A.1 User Interfaces

Figures 5 and 6 shows the user interfaces of the first and second stages of the MITI annotation experiment conducted in UpWork. The first stage is when two workers from UpWork were asked to annotate each sentence contained in the listener utterances of the dialogues from CounselChat and RED datasets and the second stage is when a high quality worker was asked to act as a judge to resolve the disagreements occured in the first stage. Interfaces similar to the second stage were used in the third stage as well. To educate the worker on the MITI coding scheme and the labels we derived out of it, a detailed tutorial was shown to the worker at the beginning of the task. This is shown in Figure 3c. A practice task to self-evaluate their competence in annotating responses with the labels derived from the MITI code followed next. Figure 4c depicts this.

A.2 More About the MITI

The MITI does not contain an exhaustive list of all possible codes; thus not all sentences can be mapped to a label from the MITI code. In this case, the annotators were asked to select *Other*. Also, the labels from the MITI code are mutually exclusive. Thus, the same sentence could not receive more than one label.

A.3 Worker Quality

In stage 1 of the annotation process, to motivate the workers to pay attention to the task, we offered to pay them a bonus of $5 for each batch of dialogues that scored an above average observed agreement with a peer worker. Out of 400 worker assignments (200 batches × 2 workers per batch), 140 of them (35%) were able to receive this bonus.

As the task progressed, those who scored higher observed agreements with the peer workers were allocated more batches to annotate.

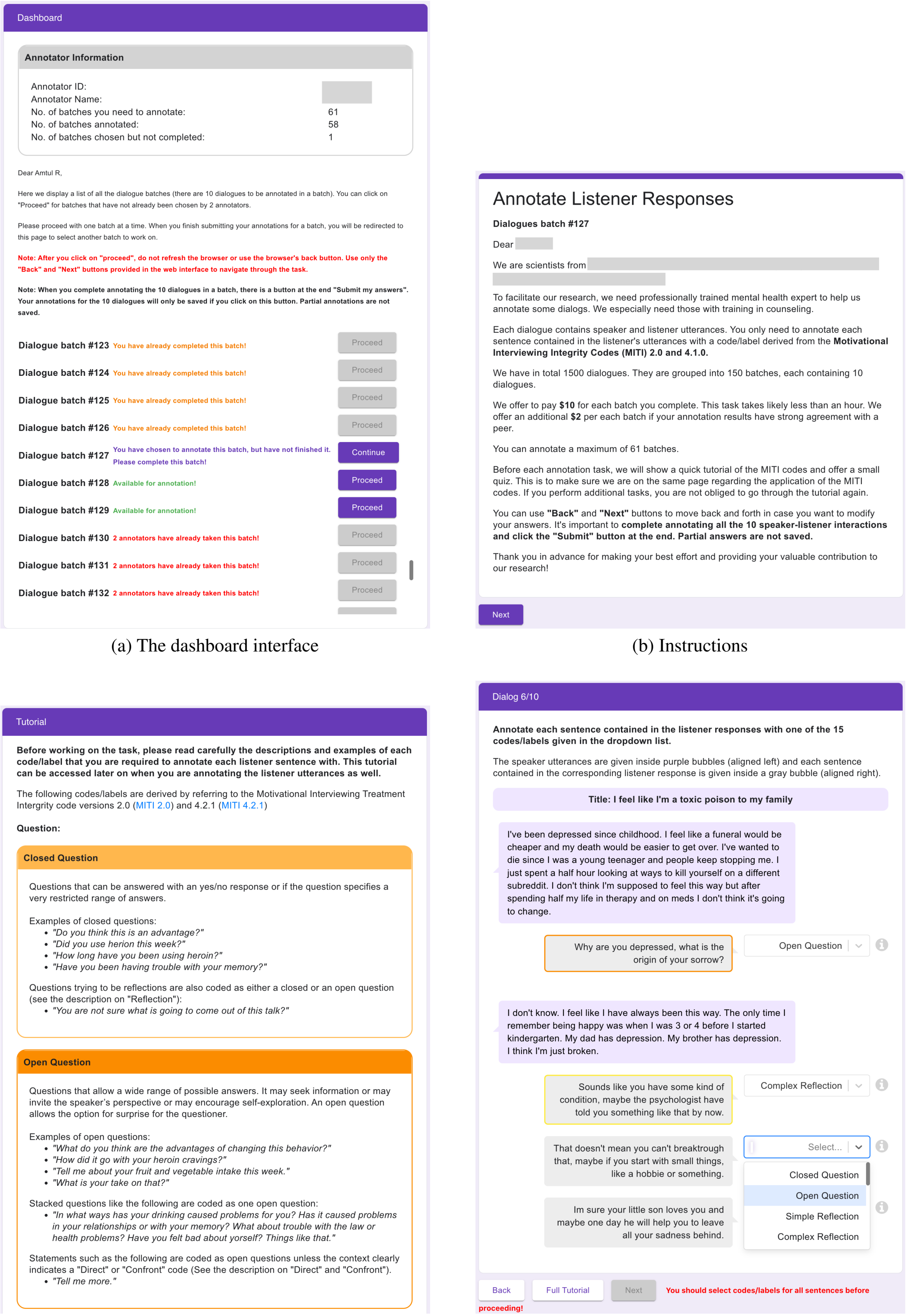
In the second and third stages, to validate the quality of the judges and their attentiveness to the task, hidden checkpoints were included to measure the workers' attentiveness to the task. These checkpoints were based on the labels agreed by the two workers in the first stage of the task. In each batch of 10 dialogues, we randomly selected 10 sentences for which a label was agreed in the first stage. For each such sentence, we randomly sampled another label out of the remaining labels and showed it along with the correct label for the judge to select from. The four judges we recruited were able to get in overall 84.3% questions correct in stage 2 of the annotation task. The scores for each of the four judges were 80.00%, 86.47%, 86.47%, and 87.50%. In the third stage, they were able to get in overall 82.93% questions correct. Their individual scores were 83.00%, 83.64%, 82.00%, and 83.00%. All the scores being above 80% in both stages indicates that they all were paying significant attention to the task.

A.4 Confusion Matrices

Figure 5 shows the confusion matrices in stage 1 of the experiment between the two annotators for the CounselChat and RED datasets separately. Labels such as *Give Information*, *Advise without Permission*, and *Closed Question* had the highest agreement between the two workers in the CounselChat dataset, whereas in RED, the highest agreed labels were *Self-Disclose*, *Give Information*, and *Support*.

Figure 6 shows the confusion matrices between the two annotators for sentence for which the label was unresolved in stage 1 and between each of these annotators and the judge in stage 2 of the annotation process. From the second and third confusion matrices corresponding to each dataset, it could seen how the judge's annotations aligned with annotations from each annotator from stage 1.

Figure 7 shows the confusion matrices between the two remaining annotators after the annotations from the poorly performed worker are removed and between each of these annotators and the second judge in stage 2 of the annotation process. Note that in the remaining two annotations, the first one comes from a relatively better performed worker from stage 1 and the second one comes from the first judge from stage 2. By observing the confusion matrices, it was noted that 73.34% times, the second judge agreed with the annotation provided by the first judge in stage 2. This further validated the quality of the judges selected.



Dashboard

Annotator Information

Annotator ID: [redacted]
Annotator Name:
No. of batches you need to annotate: 61
No. of batches annotated: 58
No. of batches chosen but not completed: 1

Dear Amtul R,

Here we display a list of all the dialogue batches (there are 10 dialogues to be annotated in a batch). You can click on "Proceed" for batches that have not already been chosen by 2 annotators.

Please proceed with one batch at a time. When you finish submitting your annotations for a batch, you will be redirected to this page to select another batch to work on.

Note: After you click on "proceed", do not refresh the browser or use the browser's back button. Use only the "Back" and "Next" buttons provided in the web interface to navigate through the task.

Note: When you complete annotating the 10 dialogues in a batch, there is a button at the end "Submit my answers". Your annotations for the 10 dialogues will only be saved if you click on this button. Partial annotations are not saved.

Dialogue batch #123 You have already completed this batch! Proceed

Dialogue batch #124 You have already completed this batch! Proceed

Dialogue batch #125 You have already completed this batch! Proceed

Dialogue batch #126 You have already completed this batch! Proceed

Dialogue batch #127 You have chosen to annotate this batch, but have not finished it. Please complete this batch! Continue

Dialogue batch #128 Available for annotation! Proceed

Dialogue batch #129 Available for annotation! Proceed

Dialogue batch #130 2 annotators have already taken this batch! Proceed

Dialogue batch #131 2 annotators have already taken this batch! Proceed

Dialogue batch #132 2 annotators have already taken this batch! Proceed

(a) The dashboard interface

Annotate Listener Responses

Dialogues batch #127

Dear [redacted]

We are scientists from [redacted]

To facilitate our research, we need professionally trained mental health expert to help us annotate some dialogs. We especially need those with training in counseling.

Each dialogue contains speaker and listener utterances. You only need to annotate each sentence contained in the listener's utterances with a code/label derived from the **Motivational Interviewing Integrity Codes (MITI) 2.0 and 4.1.0.**

We have in total 1500 dialogues. They are grouped into 150 batches, each containing 10 dialogues.

We offer to pay **$10** for each batch you complete. This task takes likely less than an hour. We offer an additional **$2** per each batch if your annotation results have strong agreement with a peer.

You can annotate a maximum of 61 batches.

Before each annotation task, we will show a quick tutorial of the MITI codes and offer a small quiz. This is to make sure we are on the same page regarding the application of the MITI codes. If you perform additional tasks, you are not obliged to go through the tutorial again.

You can use **"Back"** and **"Next"** buttons to move back and forth in case you want to modify your answers. It's important to **complete annotating all the 10 speaker-listener interactions and click the "Submit" button at the end. Partial answers are not saved.**

Thank you in advance for making your best effort and providing your valuable contribution to our research!

Next

(b) Instructions

Tutorial

Before working on the task, please read carefully the descriptions and examples of each code/label that you are required to annotate each listener sentence with. This tutorial can be accessed later on when you are annotating the listener utterances as well.

The following codes/labels are derived by referring to the Motivational Interviewing Treatment Intergrity code versions 2.0 (MITI 2.0) and 4.2.1 (MITI 4.2.1)

Question:

Closed Question

Questions that can be answered with an yes/no response or if the question specifies a very restricted range of answers.

Examples of closed questions:
- *"Do you think this is an advantage?"*
- *"Did you use herion this week?"*
- *"How long have you been using heroin?"*
- *"Have you been having trouble with your memory?"*

Questions trying to be reflections are also coded as either a closed or an open question (see the description on "Reflection"):
- *"You are not sure what is going to come out of this talk?"*

Open Question

Questions that allow a wide range of possible answers. It may seek information or may invite the speaker's perspective or may encourage self-exploration. An open question allows the option for surprise for the questioner.

Examples of open questions:
- *"What do you think are the advantages of changing this behavior?"*
- *"How did it go with your heroin cravings?"*
- *"Tell me about your fruit and vegetable intake this week."*
- *"What is your take on that?"*

Stacked questions like the following are coded as one open question:
- *"In what ways has your drinking caused problems for you? Has it caused problems in your relationships or with your memory? What about trouble with the law or health problems? Have you felt bad about yourself? Things like that."*

Statements such as the following are coded as open questions unless the context clearly indicates a "Direct" or "Confront" code (See the description on "Direct" and "Confront").
- *"Tell me more."*

(c) The tutorial

Dialog 6/10

Annotate each sentence contained in the listener responses with one of the 15 codes/labels given in the dropdown list.

The speaker utterances are given inside purple bubbles (aligned left) and each sentence contained in the corresponding listener response is given inside a gray bubble (aligned right).

Title: I feel like I'm a toxic poison to my family

I've been depressed since childhood. I feel like a funeral would be cheaper and my death would be easier to get over. I've wanted to die since I was a young teenager and people keep stopping me. I just spent a half hour looking at ways to kill yourself on a different subreddit. I don't think I'm supposed to feel this way but after spending half my life in therapy and on meds I don't think it's going to change.

Why are you depressed, what is the origin of your sorrow? — Open Question

I don't know. I feel like I have always been this way. The only time I remember being happy was when I was 3 or 4 before I started kindergarten. My dad has depression. My brother has depression. I think I'm just broken.

Sounds like you have some kind of condition, maybe the psychologist have told you something like that by now. — Complex Reflection

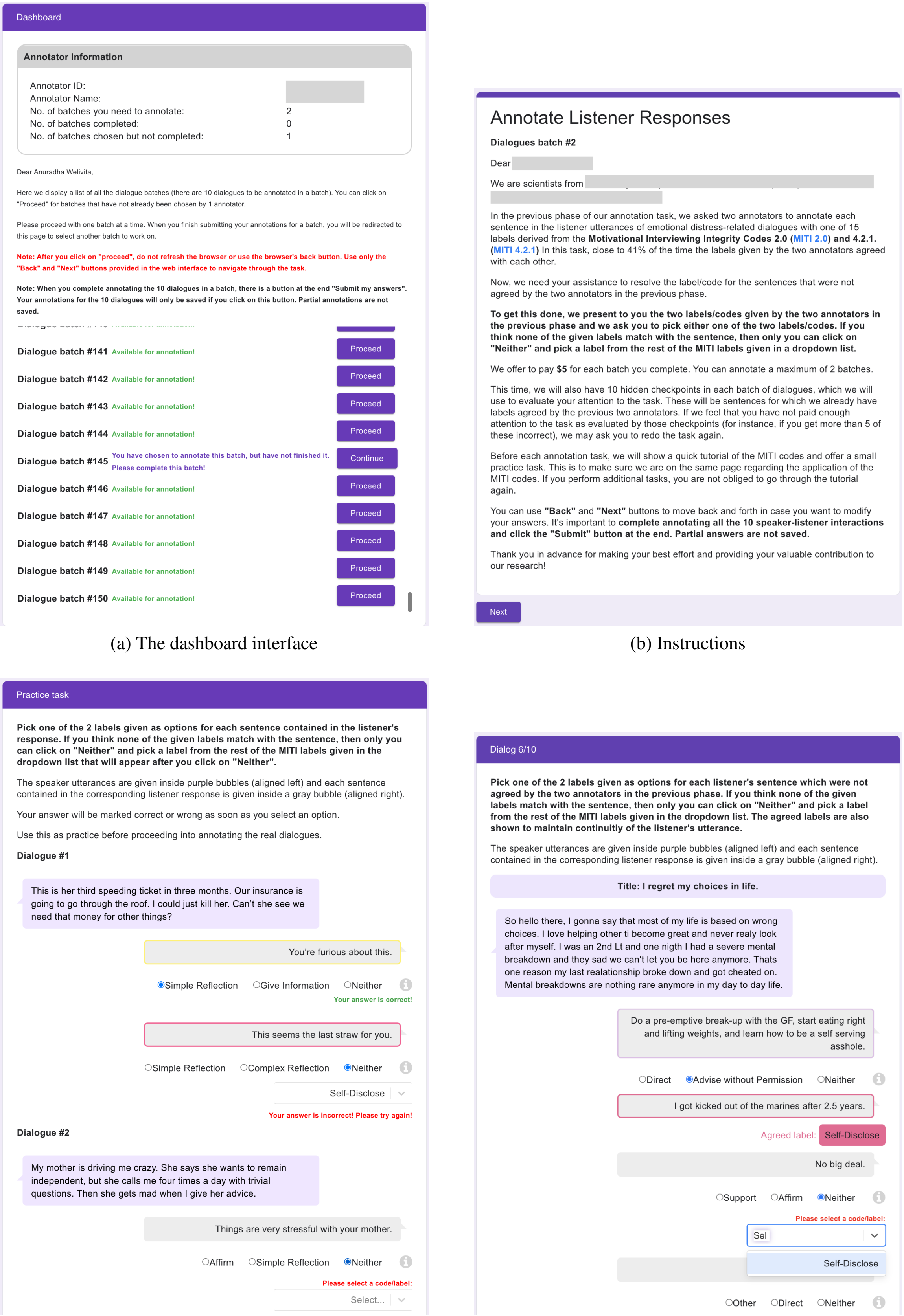
That doesn't mean you can't breaktrough that, maybe if you start with small things, like a hobbie or something. — Select...

Im sure your little son loves you and maybe one day he will help you to leave all your sadness behind.

You should select codes/labels for all sentences before proceeding!

(d) The annotation task interface

Figure 3: User interfaces of the first stage of the MITI annotation experiment.



Dashboard

Annotator Information

Annotator ID: [redacted]
Annotator Name:
No. of batches you need to annotate: 2
No. of batches completed: 0
No. of batches chosen but not completed: 1

Dear Anuradha Welivita,

Here we display a list of all the dialogue batches (there are 10 dialogues to be annotated in a batch). You can click on "Proceed" for batches that have not already been chosen by 1 annotator.

Please proceed with one batch at a time. When you finish submitting your annotations for a batch, you will be redirected to this page to select another batch to work on.

Note: After you click on "proceed", do not refresh the browser or use the browser's back button. Use only the "Back" and "Next" buttons provided in the web interface to navigate through the task.

Note: When you complete annotating the 10 dialogues in a batch, there is a button at the end "Submit my answers". Your annotations for the 10 dialogues will only be saved if you click on this button. Partial annotations are not saved.

Dialogue batch #141 Available for annotation! Proceed

Dialogue batch #142 Available for annotation! Proceed

Dialogue batch #143 Available for annotation! Proceed

Dialogue batch #144 Available for annotation! Proceed

Dialogue batch #145 You have chosen to annotate this batch, but have not finished it. Please complete this batch! Continue

Dialogue batch #146 Available for annotation! Proceed

Dialogue batch #147 Available for annotation! Proceed

Dialogue batch #148 Available for annotation! Proceed

Dialogue batch #149 Available for annotation! Proceed

Dialogue batch #150 Available for annotation! Proceed

(a) The dashboard interface

Annotate Listener Responses

Dialogues batch #2

Dear [redacted]

We are scientists from [redacted]

In the previous phase of our annotation task, we asked two annotators to annotate each sentence in the listener utterances of emotional distress-related dialogues with one of 15 labels derived from the **Motivational Interviewing Integrity Codes 2.0 (MITI 2.0) and 4.2.1. (MITI 4.2.1)** In this task, close to 41% of the time the labels given by the two annotators agreed with each other.

Now, we need your assistance to resolve the label/code for the sentences that were not agreed by the two annotators in the previous phase.

To get this done, we present to you the two labels/codes given by the two annotators in the previous phase and we ask you to pick either one of the two labels/codes. If you think none of the given labels match with the sentence, then only you can click on "Neither" and pick a label from the rest of the MITI labels given in a dropdown list.

We offer to pay **$5** for each batch you complete. You can annotate a maximum of 2 batches.

This time, we will also have 10 hidden checkpoints in each batch of dialogues, which we will use to evaluate your attention to the task. These will be sentences for which we already have labels agreed by the previous two annotators. If we feel that you have not paid enough attention to the task as evaluated by those checkpoints (for instance, if you get more than 5 of these incorrect), we may ask you to redo the task again.

Before each annotation task, we will show a quick tutorial of the MITI codes and offer a small practice task. This is to make sure we are on the same page regarding the application of the MITI codes. If you perform additional tasks, you are not obliged to go through the tutorial again.

You can use **"Back"** and **"Next"** buttons to move back and forth in case you want to modify your answers. It's important to **complete annotating all the 10 speaker-listener interactions and click the "Submit" button at the end. Partial answers are not saved.**

Thank you in advance for making your best effort and providing your valuable contribution to our research!

Next

(b) Instructions

Practice task

Pick one of the 2 labels given as options for each sentence contained in the listener's response. If you think none of the given labels match with the sentence, then only you can click on "Neither" and pick a label from the rest of the MITI labels given in the dropdown list that will appear after you click on "Neither".

The speaker utterances are given inside purple bubbles (aligned left) and each sentence contained in the corresponding listener response is given inside a gray bubble (aligned right).

Your answer will be marked correct or wrong as soon as you select an option.

Use this as practice before proceeding into annotating the real dialogues.

Dialogue #1

This is her third speeding ticket in three months. Our insurance is going to go through the roof. I could just kill her. Can't she see we need that money for other things?

You're furious about this.

● Simple Reflection ○ Give Information ○ Neither

Your answer is correct!

This seems the last straw for you.

○ Simple Reflection ○ Complex Reflection ● Neither

Self-Disclose

Your answer is incorrect! Please try again!

Dialogue #2

My mother is driving me crazy. She says she wants to remain independent, but she calls me four times a day with trivial questions. Then she gets mad when I give her advice.

Things are very stressful with your mother.

○ Affirm ○ Simple Reflection ● Neither

Please select a code/label:

Select...

Dialog 6/10

Pick one of the 2 labels given as options for each listener's sentence which were not agreed by the two annotators in the previous phase. If you think none of the given labels match with the sentence, then only you can click on "Neither" and pick a label from the rest of the MITI labels given in the dropdown list. The agreed labels are also shown to maintain continuitiy of the listener's utterance.

The speaker utterances are given inside purple bubbles (aligned left) and each sentence contained in the corresponding listener response is given inside a gray bubble (aligned right).

Title: I regret my choices in life.

So hello there, I gonna say that most of my life is based on wrong choices. I love helping other ti become great and never realy look after myself. I was an 2nd Lt and one nigth I had a severe mental breakdown and they sad we can't let you be here anymore. Thats one reason my last realationship broke down and got cheated on. Mental breakdowns are nothing rare anymore in my day to day life.

Do a pre-emptive break-up with the GF, start eating right and lifting weights, and learn how to be a self serving asshole.

○ Direct ● Advise without Permission ○ Neither

I got kicked out of the marines after 2.5 years.

Agreed label: Self-Disclose

No big deal.

○ Support ○ Affirm ● Neither

Please select a code/label:

Sel

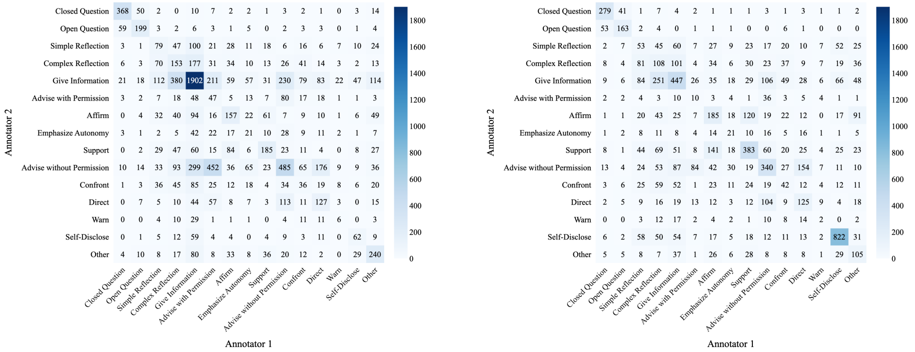
Self-Disclose

○ Other ○ Direct ○ Neither

(c) The practice task

(d) The task interface for resolving labels

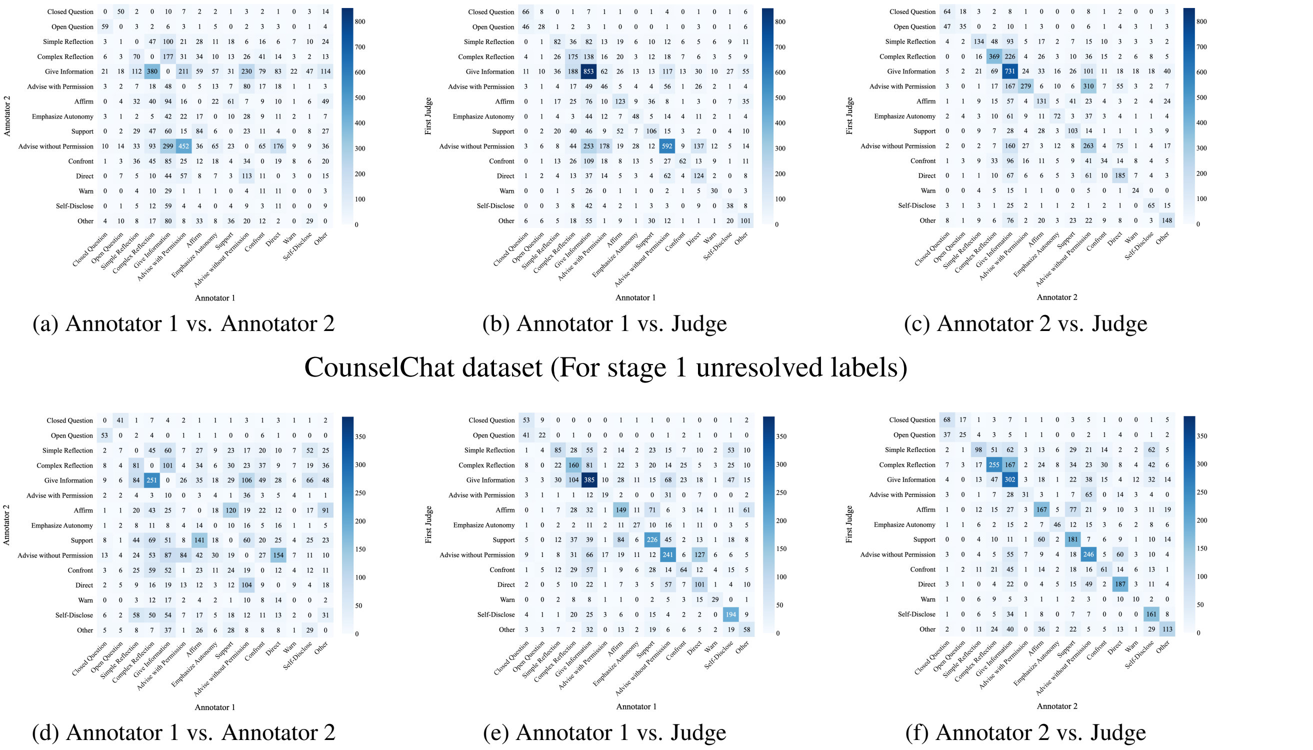
Figure 4: User interfaces of the second stage of the MITI annotation experiment.



(a) CounselChat dataset

(b) RED dataset

Figure 5: Confusion matrices between the two annotators for responses in the CounselChat and RED datasets during stage 1 of the annotation process.



(a) Annotator 1 vs. Annotator 2

(b) Annotator 1 vs. Judge

(c) Annotator 2 vs. Judge

CounselChat dataset (For stage 1 unresolved labels)

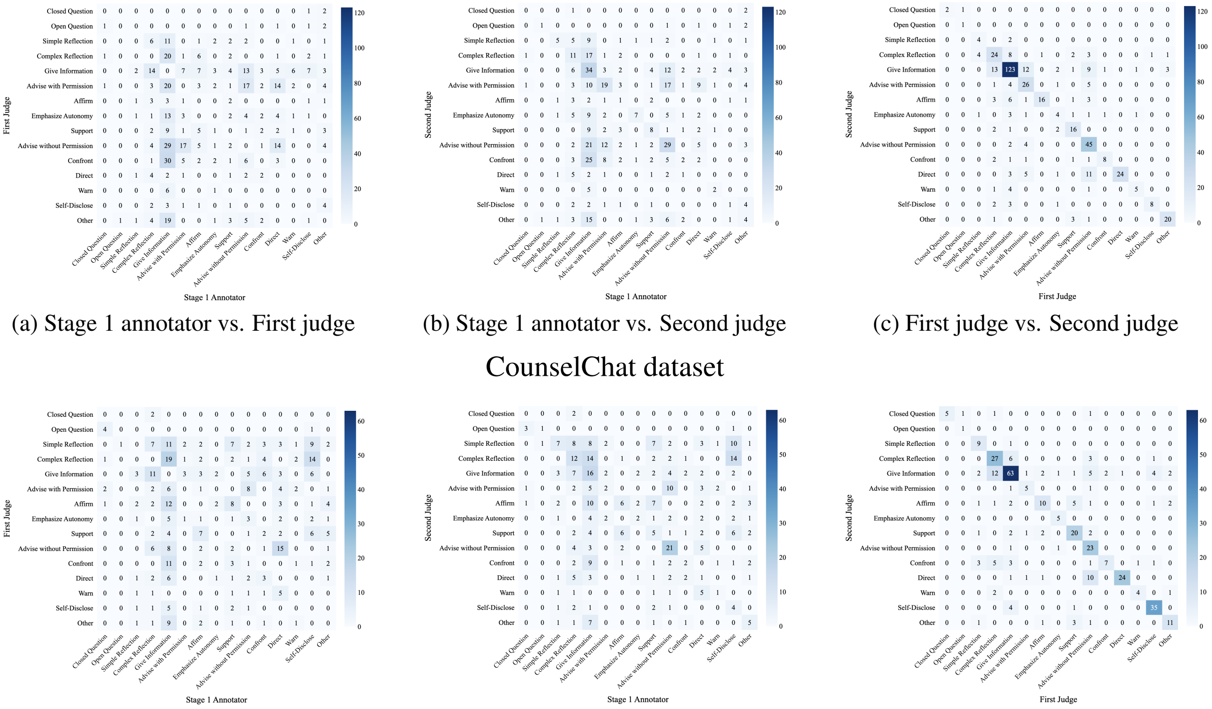
(d) Annotator 1 vs. Annotator 2

(e) Annotator 1 vs. Judge

(f) Annotator 2 vs. Judge

RED dataset (For stage 1 unresolved labels)

Figure 6: Confusion matrices between the two annotators for sentence for which the label was unresolved in stage 1 and between each of these annotators and the judge in stage 2 of the annotation process. From the second and third confusion matrices corresponding to each dataset, it could seen how the judge's annotations aligned with annotations from each annotator from stage 1.



(a) Stage 1 annotator vs. First judge

(b) Stage 1 annotator vs. Second judge

(c) First judge vs. Second judge

CounselChat dataset

(d) Stage 1 annotator vs. First judge

(e) Stage 1 annotator vs. Second judge

(f) First judge vs. Second judge

RED dataset

Figure 7: Confusion matrices between different annotators for sentences which were still unresolved after stage 2 that contained at least one annotation from a poorly performed worker. It could be observed that the second judge's annotations in stage 3 aligned mostly with the first judge's annotations in stage 2.