Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach

Ashish Sharma[♠] Inna W. Lin[♠] Adam S. Miner[♣♡] David C. Atkins[◊] Tim Althoff[♠]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♣]Department of Psychiatry and Behavioral Sciences, Stanford University
[♡]Center for Biomedical Informatics Research, Stanford University
[◊]Department of Psychiatry and Behavioral Sciences, University of Washington
{ashshar,ilin,althoff}@cs.washington.edu

ABSTRACT

Online peer-to-peer support platforms enable conversations between millions of people who seek and provide mental health support. If successful, web-based mental health conversations could improve access to treatment and reduce the global disease burden. Psychologists have repeatedly demonstrated that *empathy*, the ability to understand and feel the emotions and experiences of others, is a key component leading to positive outcomes in supportive conversations. However, recent studies have shown that highly empathic conversations are rare in online mental health platforms.

In this paper, we work towards improving empathy in online mental health support conversations. We introduce a new task of *empathic rewriting* which aims to transform low-empathy conversational posts to higher empathy. Learning such transformations is challenging and requires a deep understanding of empathy while maintaining conversation quality through text fluency and specificity to the conversational context. Here we propose PARTNER, a deep reinforcement learning (RL) agent that learns to make sentence-level edits to posts in order to increase the expressed level of empathy while maintaining conversation quality. Our RL agent leverages a policy network, based on a transformer language model adapted from GPT-2, which performs the dual task of generating candidate empathic sentences and adding those sentences at appropriate positions. During training, we reward transformations that increase empathy in posts while maintaining text fluency, context specificity, and diversity. Through a combination of automatic and human evaluation, we demonstrate that PARTNER successfully generates more empathic, specific, and diverse responses and outperforms NLP methods from related tasks such as style transfer and empathic dialogue generation. This work has direct implications for facilitating empathic conversations on web-based platforms.

ACM Reference Format:
Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, Tim Althoff. 2021. Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3450097

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3450097

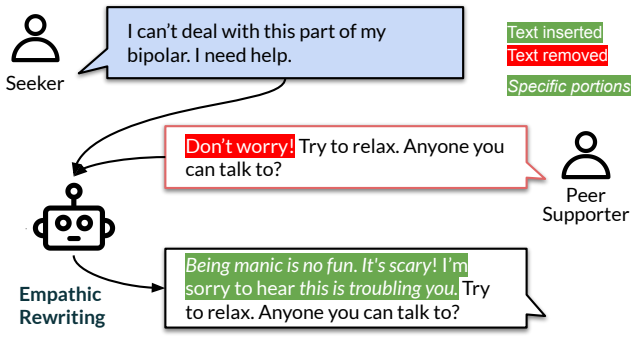


Figure 1: An overview of the empathic rewriting task. Given a post from support seeker and a low-empathy response, the task is to rewrite the response for making it more empathic, through text insertions and deletions. This task requires inferring *specific* feelings and experiences from seeker's post and using them for making appropriate changes to the response through empathic mechanisms like emotional reactions, interpretations, and explorations [59]. *Examples in this paper have been paraphrased for anonymization [42].*

1 INTRODUCTION

Online mental health support platforms such as TalkLife (talklife.co) are used by millions of users for expressing emotions, sharing stigmatized experiences, and receiving peer support. These platforms might help improve access to mental health support as mental health care remains a global challenge with widespread shortages of workforce [45], limited in-person treatment options, and other barriers like stigma [69]. A key component of providing successful support is *empathy*, the ability to understand or feel the emotions and experiences of others [17]. Quantitative evidence shows that empathic interactions have strong associations with symptom improvement in mental health support [18] and are instrumental in building therapeutic alliance and rapport [3, 54]. Yet, highly empathic conversations are rare on online support platforms [59].

Empowering peer supporters on online support platforms with feedback and training, for example through machine-in-the-loop writing systems [9, 64], has the potential to help supporters express higher levels of empathy and in turn improve the effectiveness of these platforms [26, 44, 59]. Traditional methods for training empathy (e.g., in-person counselor training) do not scale to the millions of users of online support platforms. However, computational methods that can support peer-supporters by suggesting ways to modify existing conversation utterances to make them more empathic may

help meet this need of feedback and training and indirectly benefit support seekers on the platform.

In this paper, we introduce **Empathic Rewriting**, a new task that aims to transform low-empathy conversations to higher empathy (Figure 1). For example, given a post from a support seeker "*I can't deal with this part of my bipolar. I need help.*" and a low-empathy response "*Don't worry! Try to relax. Anyone you can talk to?*", we want to increase empathy in the response by transforming it to "*Being Manic is no fun. It's scary! I'm sorry to hear this is troubling you. Try to relax. Anyone you can talk to?*"; the rewritten response should communicate more empathy through an understanding of feelings and experiences ("*Being manic is no fun. It's scary*") and display of felt emotions ("*I'm sorry to hear this is troubling you*").

Performing such transformations is a challenging task: First, empathy is a complex, conceptually nuanced construct and requires understanding the feelings and experiences shared by the support seeker. In the example above, one needs to understand that being "*bipolar*" can be "*scary*", involves "*manic*" phases, and communicate this in the response. Second, for empathic rewriting to be purposeful, it should not undermine other conversation goals like language fluency, context specificity, and diversity. Making changes that lead to ungrammatical posts with empathic portions (e.g., "*Scary it is manic being*") may not be helpful and obstruct useful feedback. Further, making the same transformation to every response (e.g., rewrite every response to "*I understand how you feel*") would lead to non-specific and generic responses reducing the overall conversational quality [30, 56]. Third, the task of empathic rewriting requires changes that go beyond simple word-level transformations, often requiring multiple new sentences to be added or replaced (e.g., three sentence insertions and one sentence removal in the example in Figure 1). This is different from related style transfer tasks [31, 61] where even changing a single word may suffice for transferring from negative to positive sentiment (e.g., replace "*bad*" with "*good*" in the sentence "*the movie was bad*"). Finally, supervised methods commonly used for similar tasks such as style transfer [31, 61] and content debiasing [39, 51] usually require a large parallel dataset. Such a dataset is not yet available for empathic rewriting and hard to collect as it would require a large number of clinical psychologists and counselors well-versed in the complex construct of empathy.

To address the challenges described above, we propose PARTNER,[1] a deep reinforcement learning (RL) model for the task of empathic rewriting (Section 5). We design an RL agent which learns to add new empathic sentences to posts or replace existing sentences in posts with more empathic ones. The agent operates on a pair of seeker post and the original response post (which rarely is highly empathic [59]) and makes edits to the response at the level of a sentence by simultaneously (a) identifying positions in the original response post where changes are required, and (b) generating empathic sentences for insertion or replacement at the identified positions (Section 5.3). We model this agent using a policy network based on a transformer decoder model adapted from GPT-2 [52]. We build upon existing large-scale pre-training of GPT-2 on conversations, as done in DialoGPT [75], and modify it to perform the two simultaneous actions of identifying positions and generating empathic sentences for empathic rewriting (Section 5.4). Through carefully constructed scoring functions, we reward transformations that increase empathy in posts while maintaining text fluency, context specificity, and diversity (Section 5.5).

Evaluating complex conversational constructs such as empathy is fundamentally challenging [59]. Therefore, we combine comprehensive automatic evaluation with expert-based human evaluation. Our experiments demonstrate that PARTNER can effectively increase empathy in posts in fluent, specific, and diverse ways and outperforms baselines used in related text generation tasks by $> 35\%$ in empathy improvement (Section 6). Also, PARTNER is the only approach that consistently improves empathy and does not lead to a *loss* of empathy when rewriting an already highly empathic post, while all baselines tend to propose a large number of edits that only make the situation worse (Section 6.1). Lastly, through comprehensive human evaluation, we show that experts in clinical psychology prefer rewritings of PARTNER compared to baselines, based on empathy, specificity, and fluency (Section 6.4). We view our approach and findings as a key step towards building AI systems for facilitating empathic conversations on online mental health support platforms, but these insights may generalize beyond mental health to other conversational settings on web-based platforms. We share our code publicly at https://github.com/behavioral-data/PARTNER.

2 RELATED WORK

We build upon prior work on NLP for online mental health support, empathic dialogue generation, reinforcement learning for text rewriting and natural language generation, and AI-assisted writing.

2.1 NLP for online mental health support

Broadly, our work relates to existing research on NLP for online mental health support. These efforts have predominantly focused on analyzing techniques that are effective for seeking and providing conversational support such as adaptability to various contexts and diversity of responses [1, 49, 60, 72, 73]. Researchers have also built methods for identifying therapeutic actions [28], quantifying language development of counselors [74], extracting patterns of conversational engagement [58], analyzing moderation [67], and detecting cognitive restructuring [50] in supportive conversations. Here, we focus on a particular conversation technique, *empathy*, which is key in counseling and mental health support [7, 17]. Our work builds on previous efforts on understanding and building computational methods for identifying empathy in online health communities [27], face-to-face therapy [20, 48], and text-based peer-to-peer support [59]. We extend this work by learning to improve empathy in online mental health support conversations through a reinforcement learning method for empathic rewriting (Section 5).

2.2 Empathic dialogue generation

Our task of empathic rewriting is related to empathic dialogue generation but has a key difference as it involves making empathic changes to *existing* responses instead of generating new responses from scratch. While research on generating empathic dialogue has mainly focused on chit-chat, open-domain conversations [34, 41, 53], we work on conversations in online mental health support. Moreover, most empathic dialogue generation methods

[1] em**PA**thic **R**ewri**T**ing in me**N**tal h**E**alth suppo**R**t

have a tendency of enabling empathic conversations through emotional grounding [53] or emotion mimicking [41]. In mental health support, however, communicating the cognitive aspects of empathy, related to understanding the experiences and feelings of others, are more valued by mental health professionals [57, 59, 65]. We extend this work with the task of empathic rewriting (Section 4) and by leveraging both emotional and cognitive aspects of empathy, using a theoretically-grounded framework of empathy [59] (Section 5).

2.3 Text rewriting and AI-assisted systems

Text rewriting is a broad subarea in natural language processing that includes tasks such as style transfer [31, 61], content debiasing [39, 51], and controllable text generation [13, 24, 40]. We propose empathic rewriting as a new text rewriting task in which conversational utterances are rewritten for increasing them in empathy (Section 4). This task presents unique challenges different from other text rewriting tasks: it requires understanding empathy in conversational contexts and leveraging that understanding for making empathic changes while ensuring high conversational quality in terms of language fluency, context specificity, and diversity.

Here, we propose a reinforcement learning (RL) model for the task of empathic rewriting (Section 5). Previous work has used RL for the task of sentiment transfer [37] by only using text generations as actions. Here, we design an RL agent that simultaneously learns to (a) identify positions for making improvements and (b) generating empathic sentences for insertion or replacement at the identified positions. These actions are important because the task of empathic rewriting requires changes that go beyond simple word-level transformations, as common in sentiment transfer tasks (e.g., change "*bland*" to "*delicious*" in "*the food was bland*" for transferring from negative to positive sentiment).

Prior work has built systems that leverage identification of effective conversational strategies such as asking open-ended questions for training users in counseling [25]. Computational methods that can perform empathic rewriting can be used for suggesting ways to make conversations more empathic in similar feedback and training systems for mental health support and counseling. In related context, researchers have built AI tools for writing assistance in negotiations [76], composing emails [8], language translation [55], creative writing [9], and communication of politeness [19].

3 DATASET DESCRIPTION

In this section, we describe the dataset used for the task of empathic rewriting.

3.1 The TalkLife platform

TalkLife (talklife.co) is the largest online peer-to-peer support platform for mental health support. It enables conversations between people seeking support (*support seekers*) and people providing support (*peer supporters*) in a thread-like setting. We call the post authored by a support seeker as *seeker post*, and the response by a peer supporter as *response post*. Table 1 describes the statistics of conversational threads on the TalkLife platform.

Curating mental health-related conversations. As noted by Sharma et al. [59], the TalkLife platform hosts a significant number of common social media interactions (e.g., *Happy mother's day*).

Dataset Statistics	TalkLife
# of Seeker posts	10.9M
# of Response posts	26.9M
# of Users	642K
Observation Period	May 2012 to June 2020

Table 1: Statistics of the TalkLife dataset.

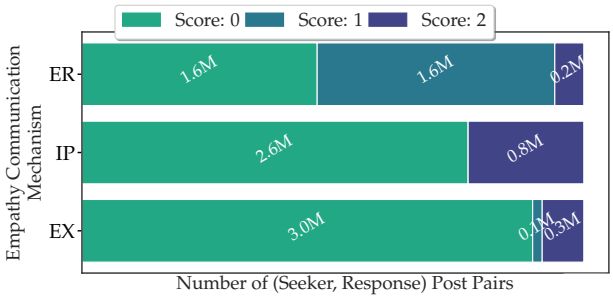


Figure 2: Expression of high levels of empathy is very low in online support platforms, especially for Interpretations (IP) and Explorations (EX). Emotional reactions (ER) are slightly more common.

Here, we focus our analyses on mental health-related conversations and filter out such posts. We manually annotate ~3k posts with answers to the question *"Is the seeker talking about a mental health related issue or situation in his/her post?"*. Using this annotated dataset, we train a standard text classifier based on BERT [15] (achieving an accuracy of ~85%). We apply this classifier to the entire TalkLife dataset and create a filtered dataset of mental health-related conversations. This dataset contains 3.33M interactions from 1.48M seeker posts.

3.2 Creating a dataset of empathic posts

Training supervised methods would require a large parallel dataset of corresponding pairs of posts with low and high empathy, respectively. As empathy is a complex phenomenon, collecting such a dataset is challenging and would likely require psychology experts. Here, we create a large non-parallel dataset with empathy measurements for training unsupervised and self-supervised computational models and a small parallel dataset with expert empathic rewritings for conducting evaluations.

Computational labeling with empathy measurements. We computationally label our dataset of 3.33M interactions with empathy measurements using a recently proposed framework of expressed empathy in mental health support [59]. This framework consists of three empathy communication mechanisms – (1) *Emotional Reactions* (expressing emotions such as warmth, compassion), (2) *Interpretations* (communicating an understanding of feelings and experiences), and (3) *Explorations* (improving understanding of the seeker by exploring feelings and experiences). For each communication mechanism, the authors design a three-point scale (0 to 2). We computationally label all pairs of (seeker post, response post) in our dataset based on this empathy scale. For this, we use a classification model (RoBERTa-based, bi-encoder attention with an accuracy of ~80%) developed by Sharma et al. [59]. Figure 2 shows the statistics which indicate that high levels of empathy expressions are uncommon in online support platforms, highlighting the need

for building systems for improving empathy (e.g., through feedback using empathic rewriting (Section 4)). We use this dataset for a supervised warm-start training in our reinforcement learning model (Section 5.6) and for training unsupervised baselines (Section 6.2).

Expert empathic rewritings. Additionally, we create a small parallel dataset of 180 pairs of corresponding low and rewritten high empathy response posts with rewritings from people having substantial expertise in empathy, mental health, and therapy (six graduate students in clinical psychology; none are co-authors). We showed them pairs of seeker and response posts and asked them to modify the response post for improving it in empathy. This expert-based dataset is designed to represent the best possible responses and we use it as ground truth for evaluation (Section 6.4).

3.3 Privacy, ethics, and disclosure

The dataset was sourced with license and consent from the TalkLife platform. All personally identifiable information (user and platform identifiers) in our dataset was removed. This work was approved by University of Washington's Institutional Review Board. We do not make any treatment recommendations or diagnostic claims.

Towards preventing unsafe rewritings. We acknowledge that building computational models for intervention in high-stakes settings such as mental health necessitates ethical considerations. There is a risk that in attempting to help, responses could have the opposite effect, which could be deadly in cases of self-harm. No current computational approach will identify and respond to harm-related utterances perfectly [43]. Thus, risk mitigation steps are appropriate in this context. Here, we remove all posts that contain a pre-defined unsafe regular expression (e.g., ∗*commit suicide*∗) from our analyses and training in collaboration with mental health professionals. Future work testing or deploying AI systems should assess safety-related risk, and also potential sources of bias (e.g., race, ethnicity, age, or gender bias in training data or models).

4 PROBLEM DEFINITION AND GOALS

In this section, we formulate the task of empathic rewriting and state the associated goals.

4.1 Empathic Rewriting

We introduce *empathic rewriting*, a new task that aims to transform low-empathy conversational posts to higher empathy. In contrast with empathic dialogue generation [34, 41, 53], where the objective is to generate empathic posts from scratch, this task requires making changes to existing posts in order to make them empathic. This is more consistent with realistic use-cases in difficult, high-stakes settings such as online support systems, which are likely to augment, rather than replace humans [44].

Formally, let S_i be a seeker post and R_i be a corresponding response post. We aim to transform R_i into its more empathic counterpart $\hat{R}_i$.

4.2 Goals

For empathic rewriting to be useful in improving mental health support conversations, the rewriting process should achieve specific goals related to empathy, conversation and natural language generation quality, and purposeful and precise feedback:

Theoretically-grounded empathy. Empathy is complex and conceptually nuanced; over time psychology research has emphasized multiple aspects of empathy [2, 4, 14, 16]. For example, computational research typically defines empathy as reacting with emotions of warmth and compassion [6]. However, psychotherapy research emphasizes aspects of empathy related to communicating cognitive understanding of feelings and experiences of others [57]. For empathic rewriting to be useful and potentially adopted in online mental health support, we need to design methods grounded in psychology and psychotherapy research. Here, we adopt the theoretically-grounded framework of empathy designed by Sharma et al. [59]. We leverage empathy measurements based on this framework as (1) reward signals in our model for empathic rewriting (Section 5.5), and (2) an automatic evaluation metric for judging improvements in empathy from various rewriting models (Section 6.3).

Context specificity and response diversity. Consider a rewriting approach that transforms every response to a generic but empathic response (e.g., "*That must have been really hard for you*"). While this approach may seem to "solve" empathic rewriting, it suffers from two key issues. First, the responses generated by this approach would lack specificity to the emotions and experiences shared in the seeker post, which is important for empathy and effective mental health support [41, 54]. Second, performing this same transformation to millions of responses on online platforms would dramatically reduce response diversity which has been shown to be important for mental health support [1] as well as in general dialogue research [30, 56].

Thus, the task of empathic rewriting interplays with other issues related to conversation and natural language generation quality and effective mental health support. Ensuring that the rewritten response is specific and diverse, along with empathic is challenging but critical for obtaining purposeful transformations. In this work, we learn rewriting actions that simultaneously achieve the goals of context specificity and response diversity using a reinforcement learning approach (Section 5.5) and we evaluate these goals using a combination of automatic and human evaluation (Section 6.3,6.4).

Text fluency and sentence coherence. In addition, only generating empathic words or phrases may not be sufficient. Without appropriate measures, the rewriting process may lead to an ungrammatical, non-fluent final response (e.g., "*Scary being is it manic*"). Also, making changes that are incoherent with the original response may not be appropriate (e.g., changing "*Sorry to hear that you lost your job. I hope you get a new job soon.*" to "*Sorry to hear that you lost your job. Congrats on your job promotion. I hope you get a new job soon.*"). In this paper, we avoid such responses with non-fluent and incoherent portions through carefully constructed reward functions (Section 5.5) and conduct both automatic and human evaluations of models on text fluency and sentence coherence (Section 6.3,6.4).

Rewriting for feedback and training. An important way in which the task of empathic rewriting can be used is for providing feedback and training to people through machine-in-the-loop writing systems [9, 64]. For humans to adopt such feedback, however, the rewriting process should make changes that are precise and specific to the original response. This means that the number of changes should be kept minimal and that the changes themselves should be suitable to the original response. For example, adding

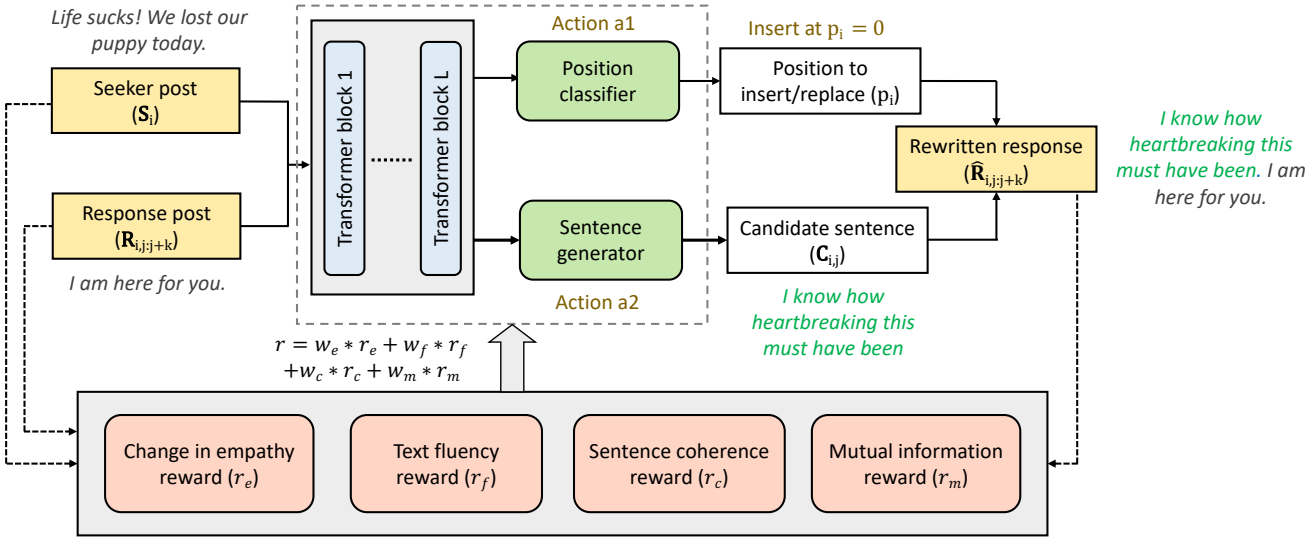


Figure 3: PARTNER uses a deep reinforcement learning approach for Empathic Rewriting. It leverages a transformer language model for performing the two actions of (1) selecting positions for insertion or replacement and (2) generating candidate empathic sentences. It uses four reward functions that promote increase in empathy, text fluency, sentence coherence, context specificity, and diversity.

10 sentences to a one-sentence response may not be useful. Here, we train a reinforcement learning agent which learns when to stop making changes through a special "stopping" action (Section 5.3). We evaluate the number of transformations different models need for empathic rewriting through a standard edit-distance based scoring metric (Section 6.3).

5 PARTNER: EMPATHIC REWRITING USING REINFORCEMENT LEARNING

Here, we present PARTNER, a reinforcement learning model for the task of empathic rewriting. We first explain the general reinforcement learning framework and its applicability to our setting. We then describe the various components of our model (states, actions, policy, and rewards) and our training strategy.

5.1 Reinforcement Learning Framework

We adopt the standard reinforcement learning framework consisting of a collection of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a policy π, and rewards $\mathcal{R}$ [63]. In this framework, given a state $s \in \mathcal{S}$, an agent takes an action $a \in \mathcal{A}$ according to the policy $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$. The policy defines whether the agent should take action a in a state s. The goal of the reinforcement learning agent is to learn a policy which maximizes the reward $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$.

Here, we design a reinforcement learning model for the task of empathic rewriting. Conceptually, our agent leverages context from the seeker post which it uses for making specific empathic changes. Alongside, it operates on the response post, looks for areas where empathy could be improved, and works on those improvements in fluent, coherent, specific, and diverse ways. Moreover, it ensures that the changes are minimal and precise by learning when to stop through a special "stopping" action.

In our reinforcement learning model, we construct states based on seeker posts and fixed-length contiguous spans in the associated response posts (Section 5.2). Insertion, replacement, and deletion of sentences in response posts are defined as actions (Section 5.3). We learn a policy that uses transformer language models at its core (Section 5.4). We design a reward function that favors empathic, fluent, coherent, specific, and diverse transformations (Section 5.5).

5.2 State: seeker post & fixed-length contiguous spans of response post

Our agent simultaneously operates on seeker post and fixed-length contiguous spans of response post. The use of seeker post helps us in leveraging conversational context, thereby enabling transformations that are specific to the feelings and experiences shared in the seeker post. The response post is used for making transformations. The use of fixed-length contiguous spans enables a static action set.

Formally, let $\mathrm{R_i}$ contain n sentences $\mathrm{R_{i,1}}, ..., \mathrm{R_{i,n}}$. At each step, we focus on a contiguous window of k sentences starting from the jth sentence $\mathrm{R_{i,j:j+k}} = \mathrm{R_{i,j}}, ..., \mathrm{R_{i,j+k-1}}$. Then, our state $s \in \mathcal{S}$ is denoted by the pair $(\mathrm{S_i}, \mathrm{R_{i,j:j+k}})$. Our policy uses a string containing $\mathrm{S_i}$ concatenated with $\mathrm{R_{i,j:j+k}}$ separated by a special <SPLIT> token (as commonly used in BERT-like models [15]).

5.3 Actions: sentence-level edits

Our agent takes actions at the level of a sentence, i.e. it either inserts new sentences or replaces existing sentences with newer ones. A deletion operation is equivalent to replacing a sentence with an empty string. Our agent can make word-level changes by replacing the original sentence with a slightly different sentence containing only word-level edits. We focus on sentence-level edits because the task of empathic rewriting requires changes that go beyond simple word-level edits. Empathic responses typically contain multiple sentences with different goals such as emotional reactions, interpretations, and explorations [59]; generating these sentences

and using them for making changes to the response is important for empathic rewriting.

In a state (S_i, $R_{i,j:j+k}$), our agent simultaneously takes two actions – (a_1) select a position in $R_{i,j:j+k}$ for insertion or replacement, (a_2) generate a candidate empathic sentence. The action space $\mathcal{A}_1$ of a_1 consists of 2k+2 actions – k+1 positions for insertions, k positions for replacements, and one *special* action for no insertion or replacement, which stops the agent from making any further changes. The action space $\mathcal{A}_2$ of a_2 consists of all arbitrary-length sentences. We denote the action taken by our agent as $a = (a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2$.

5.4 Policy

At its core, our policy has a transformer language model consisting of a stack of masked multi-head self-attention layers, based on GPT-2 (for a detailed description, see Vaswani et al. [66], Radford et al. [52]). It takes as input an encoded representation of our state (S_i, $R_{i,j:j+k}$) and generates the action $a = (a_1, a_2)$.

($a1$) Selecting a position for insertion or replacement. Given (S_i, $R_{i,j:j+k}$) as input, we want to identify a position p_i in $R_{i,j:j+k}$ where changes need to be made for improving empathy through insertion or replacement operations. A k sentence window $R_{i,j:j+k}$ has $k+1$ positions for insertions and k positions for replacement. Then, our task is to select one of these $2k+1$ positions. We formulate this as a classification problem with $2k+2$ classes. The first $2k+1$ classes represent one of the $2k+1$ potential positions and the last class represents the "stopping" action of not selecting any position, thereby stopping the agent from making any changes and keeping the response span unchanged.

For selecting this position, we first encode the input string "S_i <SPLIT> $R_{i,j:j+k}$" using the transformer block of GPT-2. We then pass this encoded representation through a linear layer to get the prediction $\hat{p}_i$ of the position for insertion or replacement. We denote our position classifier as p_{pos}.

($a2$) Generating a candidate sentence. Given (S_i, $R_{i,j:j+k}$) as input, we want to generate a candidate sentence $C_{i,j}$ to be used for making changes to $R_{i,j:j+k}$. We frame this task as a language modeling problem where the objective is to generate $C_{i,j}$ that maximizes the conditional probability $p_{sent}(C_{i,j}|S_i, R_{i,j:j+k})$.

Similar to the position selection action, we first encode our input string "S_i <SPLIT> $R_{i,j:j+k}$" using the transformer block of GPT-2. We then compute a probability distribution over vocabulary tokens by transforming the encoded representation into a vocabulary-sized vector through a softmax layer. Finally, we use top-p sampling [23][2] over this probability distribution to generate the desired $C_{i,j}$. The generation is terminated when the sampling process encounters a special end-of-sequence token.

5.5 Rewards

Our reward functions aim to increase empathy in posts and maintain text fluency, sentence coherence, context specificity, and diversity:

[2]For generating every word in a sequence, top-p sampling (or nucleus sampling) chooses from the smallest set of words whose total probability is more than p.

Change in empathy. The task of empathic rewriting requires transformations that can increase empathy of posts. Thus, we want to reward actions that increase empathy of R_i and penalize actions that decrease empathy of R_i. Let $f_e(\cdot)$ be a function that measures empathy of posts. Then, the change in empathy reward, r_e, is defined as:

$$r_e = f_e(\hat{R}_i) - f_e(R_i) \tag{1}$$

Here, we estimate $f_e(\cdot)$ using the empathy classification model developed by Sharma et al. [59] for predicting empathy levels of responses. Sharma et al. [59] leverage a theoretically-grounded framework of empathy consisting of three empathy communication mechanisms (emotional reactions, interpretations, and explorations) and devise a scale of empathy levels from 0 to 6. They train a classification model (RoBERTa [36], accuracy ~ 80%) for predicting empathy of response posts on this scale. We use their trained model as $f_e(\cdot)$ which gives us empathy scores of $\hat{R}_i$s in the range of 0 to 6.

Text fluency. We want to prevent actions that lead to outputs that are highly empathic but not fluent or grammatically correct. Therefore, we want to reward actions that lead to fluent outputs and penalize actions resulting in non-fluent outputs. Here, we operationalize *text fluency* as the inverse of perplexity of the generated $\hat{R}_i$s. We define the text fluency reward, r_f as:

$$r_f = p_{LM}\left(\hat{R}_i\right)^{(1/N)} \tag{2}$$

where p_{LM} is a general language model for English and N is the number of words in $\hat{R}_i$. Here, we use GPT-2 [52] as our p_{LM}, following previous work [12, 39].

Sentence coherence. A key component of our action space is the addition of the candidate sentence to the original response. While the candidate sentence might be highly empathic and fluent, it may not be well-suited for the response R_i to which it would be added, leading to incoherent sentences in the transformed response $\hat{R}_i$. This may not be handled by perplexity which tends to give high scores to posts where individual sentences are all fluent but are not coherent at the macro response level. Here, we design a reward function, r_c that measures coherence of the candidate sentence $C_{i,j}$ with the response span $R_{i,j:j+k}$. r_c measures the average sentence coherence probability between a candidate sentence and existing sentences in the response.

First, we create a dataset of likely coherent and incoherent sentence pairs. Given two sentences $R_{i,j1}$ and $R_{i,j2}$ in a response R_i, we call ($R_{i,j1}$, $R_{i,j2}$) a *potential coherent sentence pair*. We randomly sample a sentence R' which is not a part of responses posted to the current seeker post S_i and call (r', $R_{i,j}$) a *potential incoherent sentence pair* ($\forall R_{i,j} \in R_i$). Next, we train a text classification model, based on BERT [15], on this dataset. We take softmax at the last layer which gives us probabilities of a sentence pair being coherent ($p_{coherent}$) or incoherent ($p_{incoherent}$). Then, our sentence coherence reward is defined as:

$$r_c = \frac{\sum_{l=j}^{l=j+k} p_{coherent}\left(C_{i,j}, R_{i,l}\right)}{k} \tag{3}$$

Mutual information for specificity and diversity. In the pro-

cess of empathic rewriting, the final rewritten response may become generic (e.g., "*I understand how you feel*") thereby affecting the overall conversation quality [30, 56]. In order to ensure specificity to the seeker post and diversity of responses, we exploit the idea of maximizing mutual information between seeker post and the rewritten response post [30, 32]. Our mutual information reward is:

$$r_m = \lambda_{\text{MI}} * \log \overrightarrow{p}\left(\hat{\text{R}}_\text{i}|\text{S}_\text{i}\right) + (1-\lambda_{\text{MI}}) * \log \overleftarrow{p}\left(\text{S}_\text{i}|\hat{\text{R}}_\text{i}\right) \quad (4)$$

where $\overrightarrow{p}$ is the transformer language model used in our policy and $\overleftarrow{p}$ is an identical language model for performing the reverse task of generating seeker post from the rewritten response.

Total reward. Our total reward is $r = w_e * r_e + w_f * r_f + w_c * r_c + w_m * r_m$.

5.6 Optimization and training

Warm-start using supervised learning. We use the pre-trained weights of DialoGPT [75] for initializing our transformer language model. Next, we use a warm-start strategy using supervised learning on a parallel dataset of (low empathy, high empathy) pairs, following previous work in reinforcement learning for dialogue generation [32]. For creating this dataset, we follow the reverse process of making highly empathic responses less empathic by removing sentences that are high in empathy. Similar "reverse-engineering" strategy has also been shown to work well in other complex linguistic phenomenon like humor [68]. We first identify highly empathic sentences (with scores ≥ 2) in our dataset of empathic interactions (Section 3.2). For a seeker post S_i and response post R_i having a highly empathy sentence $\text{R}_{\text{i,j}}$, we create a dataset with (S_i <SPLIT> R_i, $\text{R}_\text{i} - \text{R}_{\text{i,j}}$) pairs.[3] We use this dataset to finetune our DialoGPT-initialized transformer language model.

REINFORCE with a baseline value for training. We use the standard REINFORCE algorithm [70] for training our agent. Our loss function is defined as:

$$J(\theta) = -(r-b) * \left(\log p_{\text{pos}}\left(a_1|\text{S}_\text{i}, \text{R}_{\text{i,j:j+k}}\right) + \log p_{\text{sent}}\left(a_2|\text{S}_\text{i}, \text{R}_{\text{i,j:j+k}}\right)\right) \quad (5)$$

where θ is our set of parameters and b is a baseline estimate of the reward (running average of previous 100 reward values) used for stabilizing training.

Experimental setup. We use a batch size of 16 and train our model for 20000 steps using a learning rate of 1e-5. We use $w_e = 1.0$, $w_f = 10.0$, $w_c = 0.1$, and $w_m = 0.1$ (selected using a grid-search approach with three values (0.1, 1.0, 10.0) for each hyperparameter). Moreover, we choose $k = 2$, p = 0.92, and $\lambda_{\text{MI}} = 0.5$. We truncate both seeker and response post to 64 tokens each.

6 EXPERIMENTS

Next, we present experiments for analyzing the performance of PARTNER on the task of empathic rewriting. We first describe automatic evaluation metrics (Section 6.1) based on the desired goals for empathic rewriting (Section 4.2), baseline approaches and ablations (Section 6.2), and demonstrate results on the automatic evaluation metrics (Section 6.3). Since evaluation using automated metrics in language generation tasks are often not robust [35], we additionally present human evaluation results from people having expertise in therapy and mental health (Section 6.4). We end with a qualitative discussion on the model's performance (Section 6.5).

6.1 Automatic evaluation metrics

We use a number of automatic metrics that are based on the goals associated with empathic rewriting (Section 4.2):

- **Change in empathy:** A key metric for successful empathic rewriting is how much the empathy has changed from the original response to the rewritten response. Similar to our reward function (Section 5.5), we measure this change using the empathy classification model developed by Sharma et al. [59]. The model computes empathy scores in the range 0 to 6 (leading to change of empathy ranging from -6 to 6).
- **Perplexity:** Similar to our text fluency reward (Section 5.5), we measure perplexity for quantifying fluency of the rewritten responses. For this, we use a pre-trained GPT-2 language model that has not been fine-tuned on our dataset, following previous work [12, 39].
- **Sentence coherence:** Since empathic rewriting requires changes at the sentence level, ensuring coherent sentences in the final rewritten response is crucial. Here, we measure sentence coherence using the scoring mechanism developed in Section 5.5.
- **Specificity:** The rewritten response should be specific to the seeker post. Following Xu et al. [71], we measure specificity using word embedding similarity between seeker post and rewritten response post (using embeddings from BERT [15]).
- **Diversity:** Since empathic rewriting has implications on millions of conversations on online mental health platforms, ensuring diversity of responses is important. Here, we measure diversity using the *distinct-1* and *distinct-2* metrics, following Li et al. [30]. The two metrics compute the number of distinct unigrams and bigrams respectively divided by the total number of tokens.
- **Edit rate:** The changes in empathic rewriting should be minimal and precise. Here, we use edit rate [62] to measure the number of changes between the original response and the rewritten response. Edit rate is defined by the Levenshtein distance between the two responses divided by the length of the original response.

6.2 Baselines and Ablations

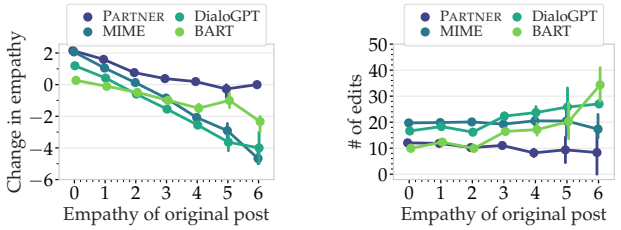
As the task of empathic rewriting has not been explored before, we compare against baseline approaches from the related tasks of dialogue generation and style transfer. Our baselines are:

- **DialoGPT [75]:** A large dialogue generation model, based on GPT-2 [52] and pre-trained on Reddit conversations.
- **MIME [41]:** An empathic dialogue generation model which exploits emotion mimicking while accounting for emotion polarity (positive or negative).
- **Deep latent sequence model [22]:** A deep generative model designed for unsupervised style transfer.
- **BART [29]:** An encoder-decoder model for sequence-to-sequence language generation.

[3] $\text{R}_\text{i} - \text{R}_{\text{i,j}}$ refers to the full response post R_i with the sentence $\text{R}_{\text{i,j}}$ removed.

	Model	Change in empathy (↑)	Perplexity (↓)	Specificity (↑)	Diversity (↑) distinct-1	Diversity (↑) distinct-2	Sentence coherence (↑)	Edit rate (↓)
Dialogue Generation	DialoGPT [75]	0.4698	8.6500	0.8921	0.0382	0.1334	0.6683	1.3520
	MIME [41]	1.2069	9.0171	0.8837	0.0031	0.0198	0.3687	1.8193
Seq-to-Seq Generation	Latent Seq. [22]	0.9745	8.7143	0.8512	0.0001	0.0002	**0.9252**	7.8853
	BART [29]	-0.0611	**7.2040**	0.8878	**0.0722**	**0.3945**	0.4560	**0.7496**
	PARTNER	**1.6410**	7.3641	**0.9052**	0.0659	0.3807	0.3030	0.9654

Table 2: Performance of PARTNER and comparisons with dialogue generation and other sequence-to-sequence generation baselines on the set of automatic metrics. PARTNER outperforms all baselines in empathy improvement and generates fluent, specific, and diverse outputs with lower edits. (↑) indicates higher is better, (↓) indicates lower is better.



(a) PARTNER and MIME are effective at increasing empathy in zero-empathy responses. However, PARTNER is more effective in increasing empathy in low, non-zero empathic responses and doesn't make an already empathic post worse.

(b) PARTNER makes lesser number of changes compared to baselines. The changes are relatively more for less empathic responses which also tend to be shorter.

Figure 4: Analysis of empathic rewritings. All error bars in this paper are 95% confidence intervals.

DialoGPT and MIME baselines completely disregard the original response; the rewritten response is the response generated given a seeker post by the respective dialogue generation models. Deep latent sequence model and BART perform a sequence-to-sequence generation from a (seeker post, original response post) pair to a response with higher empathy. We use publicly available implementations of all our baselines. We further fine-tune deep latent sequence model on the dataset of empathy-labeled interactions (Section 3.2) and BART on the heuristic-based dataset created for warm-start (Section 5.6).

Additionally, we investigate the importance of different components of our model using the following ablated baselines:

- **Warm-start only, no RL training:** We analyze the performance of the model at the end of our warm-start stage, i.e. without any RL training.
- **No coherence reward:** We train the model without using the sentence coherence reward.
- **No mutual information**: We train the model without using the mutual information component.

6.3 Automatic metrics results

Baseline Results. Table 2 reports the results of PARTNER on the automatic evaluation metrics and comparisons with baselines. We find that empathic rewriting through PARTNER achieves the largest change in empathy (35% more than the next best approach, MIME) and is more specific than all baselines. MIME generates empathic outputs (+1.21 change in empathy) but the generations have low diversity (86% less than PARTNER) indicating similar responses for most seeker posts. BART generates outputs with lowest perplexity, highest diversity, and lowest edit rate, which is consistent with substantial improvements to language models in recent years [5]. However, to our surprise, the rewritten responses through BART receive an overall drop of 0.06 in empathy, indicating that the model is unable to perform the task of empathic rewriting well and only generates non-empathic, fluent, diverse text.

Our specificity metric can be hard to interpret with values having a really small range (0.85 to 0.9). However, with human-based evaluation (Section 6.4), we find that a difference of 0.05 on this metric (between PARTNER and latent seq.) translates to a 90% preference towards PARTNER. Moreover, while PARTNER has the lowest sentence coherence score, we find that this is likely due to higher number of sentences generated by it compared to baselines. The baselines generate 1-2 sentence responses on an average, where achieving high coherence between sentences is expected (e.g., a one-sentence response by design has a coherence of 1.0). PARTNER, on the contrary, generates responses with ∼70% more sentences than baselines, affecting the overall coherence score.

Adaptability of rewritings to original post. Adapting to different types of original responses and making appropriate changes is an important aspect of empathic rewriting. A low empathic response needs a lot more improvements and edits than a highly empathic response. Figure 4a shows the change in empathy of responses given their original empathy levels. We find that PARTNER performs better than baselines in improving responses with low empathy. Importantly, only PARTNER succeeds at not deteriorating responses that are already highly empathic, indicating the effectiveness of PARTNER at adapting to responses with different empathy levels. We also analyze the number of edits by each model on responses with different original empathy levels (Figure 4b). PARTNER not only effects a greater change in empathy than baselines, it achieves so with the least number of edits for both low and high empathy responses.

Ablation Results. Table 3 reports results on ablated versions of PARTNER. Only using warm-start and no RL training is +0.2783 points better than the related off-the-shelf DialoGPT baseline on empathy improvement. However, the RL training in PARTNER further improves over this warm-start model by +0.8929 points. Using the coherence and mutual information rewards leads to small performance improvements, particularly in empathy (+0.03).

Model	Change in empathy (↑)	Perplexity (↓)	Specificity (↑)	Diversity (↑) distinct-1	Diversity (↑) distinct-2	Sentence coherence (↑)	Edit rate (↓)
PARTNER	**1.6410**	7.3641	0.9052	0.0659	0.3807	0.3030	**0.9654**
- no coherence	1.6127	7.2806	**0.9055**	0.0663	0.3844	0.3005	1.0108
- no mutual info.	1.6132	7.3274	0.9045	0.0674	0.3859	**0.3078**	1.0071
- warm-start only	0.7481	**7.1858**	0.9027	**0.0816**	**0.4238**	0.2935	1.0327

Table 3: Ablation results. Warm-start improves over DialoGPT but is still much worse than PARTNER in empathy improvement, highlighting the effectiveness of our RL-based training.

6.4 Human evaluation results

Since automatic evaluation in language generation is often not robust [35], we perform a human evaluation on our key metrics (empathy, fluency, and specificity) through A/B testing. We recruit six graduate students in clinical psychology with expertise in empathy and mental health support[4] and ask them to compare outputs from PARTNER against other baseline models, ablations, and expert empathic rewritings (Section 3.2) given the same input. Presenting a seeker post, a rewritten response post from PARTNER, and a rewritten response post from a baseline/ablation/expert-rewrite, we ask them to choose (a) response post which is more empathic, (b) response post which is more fluent, and (c) response post which is more specific. For each model, we collect evaluations on 50-100 examples.

Results: Baselines and ablations. Figure 5 shows the percentage of instances in which PARTNER was preferred over other baselines and ablations (values $> 50\%$ indicate preference towards PARTNER). We find that rewritten responses from PARTNER are preferred for empathic and specific responses over all baselines. DialoGPT is judged more fluent (Figure 4a) but generates responses following similar templates (e.g., "*I'm sorry you.... I hope you....*"). Moreover, PARTNER has ~55% preference for empathy over ablations where coherence and mutual information rewards are not used ($p < 0.01$).

Results: Expert rewritings. The most appropriate way of performing empathic rewriting is through human experts. However, experts with training in therapy and mental health support are limited [45] which makes it infeasible to employ them for millions of conversations on online support platforms. We use the small dataset of 180 empathic rewritings from experts to establish what the gold-standard performance for empathic rewritings in mental health support looks like. Unsurprisingly, experts are preferred ~80-90% times over PARTNER in empathy, fluency, and specificity ($p < 0.001$). However, in 10-20% cases PARTNER rewritings are preferred; these are typically instances where PARTNER is able to make empathic changes to responses while the experts leave it unchanged.

Results: BLEU scores. We also use the dataset of expert empathic rewritings (Section 3.2) as a ground truth of empathic rewritings and compare outputs of PARTNER, baselines, and ablations based on this ground truth using the BLEU metric [47] (Table 4). We find that the outputs from PARTNER are closest to expert rewritings (86% better than the next best baseline, BART).

[4]Most participants were PhD students in second or subsequent years of their degree program. Research in Psychology has shown that clinical psychology graduate students are, in general, representative of mental health professionals [46]. Although there are likely some differences between students and licensed psychologists, clinical outcomes in empathy-related measures such as therapeutic alliance have been shown to be comparable while students receive supervision [21].

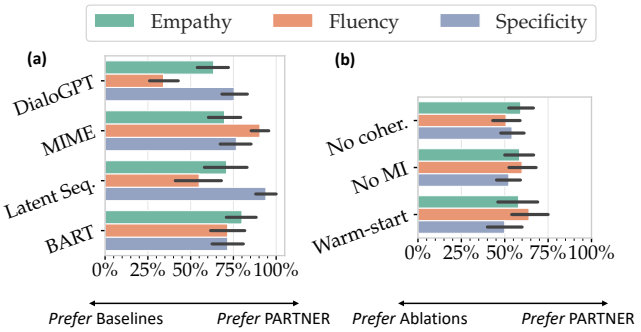


Figure 5: Human evaluation of empathy, fluency, and specificity in rewritings from PARTNER vs. (a) rewritings from baseline models, and (b) rewritings from ablations. PARTNER is preferred over baselines and ablations in empathy and specificity and is competitive in fluency.

	Model	BLEU score
	PARTNER	**0.1391**
Baselines	DialoGPT	0.0722
	MIME	0.0808
	Latent Seq	0.0254
	BART	0.0956
Ablations	- no coherence	0.1335
	- no mutual info.	0.1297
	- warm-start only	0.1281

Table 4: BLEU scores of all models against empathic rewritings from experts. The rewritings from PARTNER are closest to expert rewritings.

6.5 Qualitative examples

We present example rewritings from PARTNER and baselines in Table 5. PARTNER generates rewritings that leverage both seeker post and original response post for empathic responses. For example, from the seeker post "*I feel like nobody cares about my existence*", PARTNER is able to infer "*It's hard to find others who can relate*". Also, PARTNER can coherently transform the response post "*What happened between you two?*" to "*What happened between you two? What caused you to break?*".

Seeker post	Response post	Model	Rewritten response post
I feel like nobody cares about my existence.	I feel the same.	PARTNER	It's hard to find others who can relate. I feel the same.
		DialoGPT	I care.
		MIME	That is awful, i am sorry to hear that. I hope you feel better.
		Latent Seq.	I'm sorry to hear that.
		BART	I think i feel the same way. I do.
I overthink everything my friends say. I disregard them when they praise me but if they say something negative I take it so personally. I'm sick of being so paranoid.	I understand completely. I feel the same way.	PARTNER	It's hard when you're with such people and they are being like that to you. I understand completely. I feel the same way.
		DialoGPT	I'm the same way. I'm not sure why, but I just don't like being told I'm being paranoid.
		MIME	That is a shame, i am sure it will not be a bit of a bit!
		Latent Seq.	I know that, i feel the same way i'm sorry to.
		BART	I'm sorry, I understand completely. I feel the same way. I'm here for you.
I hope I didn't lose my only two best friends. They are the only ones I could relate to.	What happened between you two?	PARTNER	That's awful. What happened between you two? What caused you to break?
		DialoGPT	I hope you didn't lose them. I don't want to lose them.
		MIME	That is a good thing to do.
		Latent Seq.	I'm sorry to hear that.
		BART	What happened? What happened between you two?

Table 5: Qualitative examples of empathic rewriting using PARTNER and baseline methods. PARTNER generates rewritings that leverage both seeker post and original response post for empathic responses. It infers perspective from seeker posts ("*I feel like nobody cares about my existence*" → "*It's hard to find others who can relate*") and generates empathic sentences which can be coherently inserted to response posts ("*What happened between you two?*" → "*What happened between you two? What caused you to break?*")

7 DISCUSSION AND CONCLUSION

The burden of mental illness globally is overwhelming, and common mental disorders are some of the most debilitating illnesses worldwide [11]. Existing mental health resources and interventions are ill-suited to the size of the need. Online mental health support platforms that make use of peer supporters is one route to scaling up support, but the biggest challenge is to effectively train or scaffold the peer supporters. Our empathic rewriting approach represents a foundational proof-of-concept of how computational methods may help peer supporters online.

Rewriting human-generated responses may be an effective approach to balancing the benefits and risks of using artificial intelligence in mental health settings. By combining human knowledge of context and experience, our approach can both provide feedback to online peer-supporters with actionable, real-time examples, and provide support seekers with more empathic responses. Importantly, this machine-in-the-loop approach can help mitigate some of the risks related to toxicity and safety of AI systems in settings of suicidal ideation, self-harm, or insensitive comments related to race/ethnicity/gender [10, 33, 38].

Summary of contributions. Our work proposes a new task of empathic rewriting for transforming low-empathy conversational posts in online mental health support platforms to higher empathy. For this task, we develop and train PARTNER, a reinforcement learning model which makes sentence-level edits to posts for making them empathic. Through extensive experiments based on automatic and human evaluation, we show that PARTNER can effectively generate more empathic posts and outperforms baseline methods from related tasks.

ACKNOWLEDGMENTS

We would like to thank TalkLife and Jamie Druitt for their support and for providing us access to a TalkLife dataset. We also thank the members of UW Behavioral Data Science Group and the anonymous

reviewers for their suggestions and feedback. This research has been supported in part by a Microsoft AI for Accessibility grant, the Allen Institute for Artificial Intelligence, NSF grant IIS-1901386, and Bill & Melinda Gates Foundation (INV-004841). A.S.M. was supported by grants from the National Institutes of Health, National Center for Advancing Translational Science, Clinical and Translational Science Award (KL2TR001083 and UL1TR001085) and the Stanford Human-Centered AI Institute. D.C.A. was supported in part by an NIAAA K award (K02 AA023814).

Conflict of Interest Disclosure. D.C.A. is a co-founder with equity stake in a technology company, Lyssn.io, focused on tools to support training, supervision, and quality assurance of psychotherapy and counseling.

REFERENCES

[1] Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *TACL* (2016).

[2] C Daniel Batson. 2009. These things called empathy: eight related but distinct phenomena. (2009).

[3] Arthur C Bohart, Robert Elliott, Leslie S Greenberg, and Jeanne C Watson. 2002. Empathy. *J. C. Norcross (Ed.), Psychotherapy relationships that work: Therapist contributions and responsiveness to patients* (2002).

[4] Arthur C Bohart and Leslie S Greenberg. 1997. *Empathy reconsidered: New directions in psychotherapy.* American Psychological Association.

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[6] Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling Empathy and Distress in Reaction to News Stories. In *EMNLP*.

[7] Louis G Castonguay and Clara E Hill. 2017. *How and why are some therapists better than others?: Understanding therapist effects.* American Psychological Association.

[8] Mia Xu Chen, Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M Dai, Zhifeng Chen, et al. 2019. Gmail smart compose: Real-time assisted writing. In *SIGKDD*.

[9] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *IUI*.

[10] Sunny Collings and Thomas Niederkrotenthaler. 2012. Suicide prevention and emergent media: surfing the opportunity.

[11] Pamela Y Collins, Vikram Patel, Sarah S Joestl, Dana March, Thomas R Insel, Abdallah S Daar, Isabel A Bordin, E Jane Costello, Maureen Durkin, Christopher Fairburn, et al. 2011. Grand challenges in global mental health. *Nature* (2011).

[12] Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. ACL.

[13] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *ICLR*.

[14] Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy. *Journal of Personality and Social Psychology* (1980).

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

[16] Changming Duan and Clara E Hill. 1996. The current state of empathy research. *Journal of counseling psychology* (1996).

[17] Robert Elliott, Arthur C Bohart, Jeanne C Watson, and Leslie S Greenberg. 2011. Empathy. *Psychotherapy* (2011).

[18] Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy* (2018).

[19] Liye Fu, Susan R. Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. Facilitating the Communication of Politeness through Fine-Grained Paraphrasing. In *EMNLP*.

[20] James Gibson, Doğan Can, Bo Xiao, Zac E Imel, David C Atkins, Panayiotis Georgiou, and Shrikanth S Narayanan. 2016. A Deep Learning Approach to Modeling Empathy in Addiction Counseling. *Interspeech* (2016).

[21] Lizabeth A Goldstein, Abby D Adler Mandel, Robert J DeRubeis, and Daniel R Strunk. 2020. Outcomes, skill acquisition, and the alliance: Similarities and differences between clinical trial and student therapists. *Behaviour research and therapy* (2020).

[22] Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. A Probabilistic Formulation of Unsupervised Text Style Transfer. In *ICLR*.

[23] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

[24] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *ICML*.

[25] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in Building Intelligent Open-domain Dialog Systems. *ACM Transactions on Information Systems (TOIS)* (2020).

[26] Zac E Imel, Mark Steyvers, and David C Atkins. 2015. Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy* (2015).

[27] Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *IJCNLP*.

[28] Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathleen McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*.

[29] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[30] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL-HLT*.

[31] Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer. In *ACL*.

[32] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *EMNLP*.

[33] Ron C Li, Steven M Asch, and Nigam H Shah. 2020. Developing a delivery science for artificial intelligence in healthcare. *NPJ Digital Medicine* (2020).

[34] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *EMNLP*.

[35] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*.

[36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[37] Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *IJCAI*.

[38] David D Luxton, Jennifer D June, and Jonathan M Fairall. 2012. Social media and suicide: a public health perspective. *American journal of public health* (2012).

[39] Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. EMNLP.

[40] Florian Mai, Nikolaos Pappas, Ivan Montero, Noah A Smith, and James Henderson. 2020. Plug and Play Autoencoders for Conditional Text Generation. In *EMNLP*.

[41] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking Emotions for Empathetic Response Generation. In *EMNLP*.

[42] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. 2017. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *CHI*.

[43] Adam S Miner, Albert Haque, Jason A Fries, Scott L Fleming, Denise E Wilfley, G Terence Wilson, Arnold Milstein, Dan Jurafsky, Bruce A Arnow, W Stewart Agras, et al. 2020. Assessing the accuracy of automatic speech recognition for psychotherapy. *NPJ Digital Medicine* (2020).

[44] Adam S Miner, Nigam Shah, Kim D Bullock, Bruce A Arnow, Jeremy Bailenson, and Jeff Hancock. 2019. Key considerations for incorporating conversational AI in psychotherapy. *Frontiers in psychiatry* 10 (2019).

[45] Mark Olfson. 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Affairs* (2016).

[46] Lars-Göran Öst, Anna Karlstedt, and Sara Widén. 2012. The effects of cognitive behavior therapy delivered by students in a psychologist training program: An effectiveness study. *Behavior Therapy* (2012).

[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

[48] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *ACL*.

[49] 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 *ACL*.

[50] Yada Pruksachatkun, Sachin R Pendse, and Amit Sharma. 2019. Moments of Change: Analyzing Peer-Based Cognitive Support in Online Mental Health Forums. In *CHI*.

[51] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *AAAI*.

[52] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[53] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *ACL*.

[54] Elliot Robert, Arthur C Bohart, JC Watson, and LS Greenberg. 2011. Empathy. *Psychotherapy* (2011).

[55] Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. INMT: Interactive Neural Machine Translation Prediction. In *EMNLP (System Demonstrations)*.

[56] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *NAACL-HLT*.

[57] Robert L Selman. 1980. *Growth of interpersonal understanding*. Academic Press.

[58] Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020. Engagement Patterns of Peer-to-Peer Interactions on Mental Health Platforms. In *ICWSM*.

[59] 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. In *EMNLP*.

[60] Eva Sharma and Munmun De Choudhury. 2018. Mental Health Support and its Relationship to Linguistic Accommodation in Online Communities. In *CHI*.

[61] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *NeurIPS*.

[62] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, Vol. 200. Cambridge, MA.

[63] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[64] Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. Development and Evaluation of ClientBot: Patient-Like Conversational Agent to Train Basic Counseling Skills. *JMIR* (2019).

[65] CB Truax and RR Carkhuff. 1967. Modern applications in psychology. *Toward effective counseling and psychotherapy: Training and practice. Hawthorne, NY, US: Aldine Publishing Co* (1967).

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

[67] David Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The Effect of Moderation on Online Mental Health Conversations. In *ICWSM*.

[68] Robert West and Eric Horvitz. 2019. Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances". In *AAAI*.

[69] Marsha White and Steve M Dorman. 2001. Receiving social support online: implications for health education. *Health education research* (2001).

[70] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* (1992).

[71] Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. ACL.

[72] Diyi Yang, Zheng Yao, Joseph Seering, and Robert Kraut. 2019. The Channel Matters: Self-disclosure, Reciprocity and Social Support in Online Cancer Support Groups. In *CHI*.

[73] Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards. In *ACL*.

[74] Justine Zhang, Robert Filbin, Christine Morrison, Jaclyn Weiser, and Cristian Danescu-Niculescu-Mizil. 2019. Finding Your Voice: The Linguistic Development of Mental Health Counselors. In *ACL*.

[75] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *ACL, system demonstration*.

[76] Yiheng Zhou, He He, Alan W Black, and Yulia Tsvetkov. 2019. A Dynamic Strategy Coach for Effective Negotiation. In *SIGdial*.