Automatic Reflection Generation for Peer-to-Peer Counseling

Emma O'Neil
University of Pennsylvania
emoneil@alumni.upenn.edu

João Sedoc
New York University
jsedoc@stern.nyu.edu

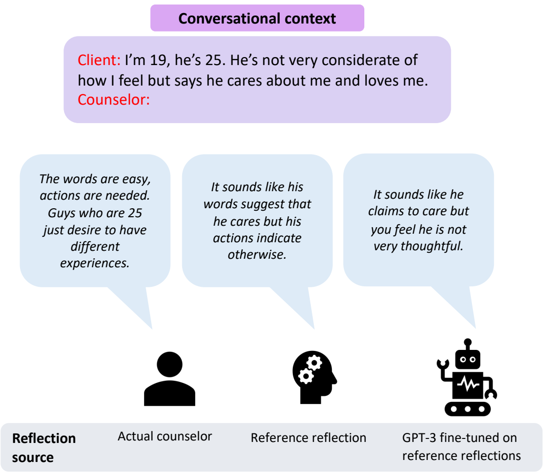
Diyi Yang
Stanford University
diyiy@cs.stanford.edu

Haiyi Zhu
Carnegie Mellon University
haiyiz@cs.cmu.edu

Lyle Ungar
University of Pennsylvania
ungar@cis.upenn.edu

Abstract

Online peer counseling platforms enable conversations between millions of people seeking and offering mental health support. Among counseling skills, reflective listening, i.e., capturing and returning to the client something the client has said, is important for positive therapeutic outcomes. We introduce a reflection generation system for online mental health support conversations leveraging GPT-3, a large language model. We compare few-shot learning against fine-tuning and assess the impact of the quality of training examples as measured by fluency, reflection resemblance, and overall preference. Fine-tuned GPT-3 generates responses that human evaluators rate as comparable in reflection quality to responses used for tuning. Models based on high-quality responses generate substantially better reflections than ones tuned on actual responses from a large online counseling service–and better reflections than the actual counselor responses. These results suggest the care needed in selecting examples for tuning generative models.



Client: I'm 19, he's 25. He's not very considerate of how I feel but says he cares about me and loves me.
Counselor:

The words are easy, actions are needed. Guys who are 25 just desire to have different experiences.

It sounds like his words suggest that he cares but his actions indicate otherwise.

It sounds like he claims to care but you feel he is not very thoughtful.

Figure 1: Illustration of conversational context and reflection response from actual counselor, reference reflection, and response by GPT-3 model fine-tuned on 350 reference reflections.

1 Introduction

Online mental health support platforms, from Talkspace to 7 Cups to Crisis Text Line, are used by millions of users for expressing challenges and receiving peer support. These platforms can help improve access to mental health support, as such care remains a global challenge with workforce shortages and limited affordable options (Olfson, 2016). Helping counselors with feedback, suggestions, and training, for instance through machine-in-the-loop writing systems (Tanana et al., 2019; Clark et al., 2018), has the potential to aid counselors in improving the quality of their responses and in turn improve the effectiveness of these platforms (Imel et al., 2015; Miner et al., 2019).

Moreover, training counselors can require substantial time and effort. Often, training incorporates didactic instruction and experiential exercises (e.g., role-playing, standardized patients, or practice with real clients) (Madson et al., 2009). In counselor training, which incorporates development of empathy and reflective listening, feedback and coaching can notably improve counselor skills (Miller et al., 2004). But standard methods for providing systematic feedback do not scale (Atkins et al., 2014). With the millions of users of online support platforms, computational methods can help counselors by suggesting draft utterances, providing guidance that may help meet the need for feedback and indirectly benefit people reaching out for support.

Among counseling skills, reflective listening is an important skill for positive therapeutic outcomes (Moyers et al., 2009). Reflective listening is one of the best ways to express to clients that they are understood (Csillik, 2013; Miller and Rollnick, 2012). Reflections capture and return to the client something the client has said. The reflected content is usually, but not always, in the client's immediately preceding utterance. For example, consider the

Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 62–75
December 6, 2023 ©2023 Association for Computational Linguistics

following client utterance and counselor reflection:

> Client: I feel so anxious when I wake up in the morning that I can't resist having some alcohol before I leave for work.
> Counselor: It sounds like it has been a way of calming your anxiety at the start of the day and releasing the tension you are feeling.

Motivated by the importance of reflective listening in counseling interactions and the value of suggestions for counselor effectiveness, we introduce a reflection generation system leveraging GPT-3, a large language model (see Figure 1). Such a system can potentially aid minimally-trained counselors in creating reflections and boost efficiency (e.g., save typing) so they have more time and can help more people. Prior work has explored the importance of ground-truth labels for in-context learning (Kim et al., 2022). Additionally, few-shot learning and fine-tuning have been examined for dialogue generation conditional on predefined stories (Miyazaki, 2023). In this work, we tackle the practical question of the impact of the quality of fine-tuning set examples on the quality of generated reflections as measured by fluency, reflection resemblance, and overall preference. We also address technical questions regarding the impact of few-shot learning versus fine-tuning and the number of examples for fine-tuning in this peer-to-peer counseling context, which have implications for model development costs. We do not find evidence to suggest fine-tuning GPT-3 on hundreds of examples results in better quality reflections than conditioning GPT-3 with 17 examples. Human evaluations show fine-tuning GPT-3 with higher-quality examples yields more highly rated reflections, aligning with an observed difference between ratings of reference reflections and actual peer counselor reflections. These results suggest care is needed in selecting examples for tuning generative models.

2 Related Work

There is a growing body of related work aimed at building automatic tools in the form of dialogue systems for online mental health support. One significant line of work has focused on delivery of psychotherapy interventions for which conversational agents serve as counselors. Han et al. (2013) created a system recognizing what users say, predicting conversation context, and following users' feelings to generate responses based on templates designed for three counseling techniques (paraphrasing, asking open questions, and reflecting feelings). Han et al. (2015) presented a similar system, generating appropriate responses using templates by identifying user emotion and intention and extracting entities and related information from a knowledge base. In contrast to these works, our system is intended to serve as a resource for counselors in creating responses to client messages.

In line with our goal of augmenting counselors' everyday practice, related work has built technology for counselor training and feedback. Such work has explored the creation of a dialogue observer that categorizes therapist and client motivational interviewing behavioral codes and forecasts codes for upcoming utterances to help guide conversation (Cao et al., 2019). Other systems have used machine learning-based feedback for training with artificial standardized clients, providing real-time suggestions on skills to use (Tanana et al., 2019). Although they aid in the evaluation of counselor language, such tools are limited in providing easily implementable suggestions. Our goal is to present the counselor with an actionable suggestion at a particular point in the conversation.

Generative applications in peer-to-peer counseling include empathic rewriting (Sharma et al., 2021) and motivational interviewing response rephrasing (Welivita and Pu, 2023), i.e., making sentence-level edits to increase empathy while maintaining conversation quality or increase motivational interviewing adherence respectively; an AI-based tool to empower peer counselors through automatic suggestion generation for a range of counseling strategies (Hsu et al., 2023); and related work in generating reflections using GPT-2 and motivational interviewing conversations (Shen et al., 2020). The latter reflects the most relevant work. Shen et al. (2020) evaluated GPT-2 models' ability to generate reflections given dialogue history, exploring how augmenting input with reflections from similar conversations and content expansion impact quality of generated reflections. They found the GPT-2 models perform better than the baseline seq2seq model. Based on automated metrics, models with context augmentation outperform the fine-tuned GPT-2 model; however, while the systems perform on-par or above reference reflections (ground truth), there appears little difference between the GPT-2 model and models leveraging additional context expansion strategies. Shen et al. (2022) addresses a similar task but enhances generation by infusing commonsense and domain-specific knowledge. Expanding on these works, we explore additional techniques including few-shot learning and evaluate the impact of the quality of examples for fine-tuning.

3 Methods

3.1 Data

Our dataset derives from conversations between clients and counselors on a large peer-to-peer online counseling service. We have a total of 1061 observations across the training and testing datasets, with 50 additional randomly sampled examples used in defining the few-shot learning prompt or for validation purposes in tuning hyperparameters, totaling 1111 observations. These observations were sourced from a larger dataset consisting of annotations of several clinical counseling skills. Messages were annotated at utterance level with counselor verbal behaviors using the Motivational Interviewing Treatment Integrity 4.2 (MITI) and the Motivational Interviewing Skill Code 2.5 (MISC) manuals. Our focus is on counselor reflections. Our training dataset consists of a total of 911 observations, which each consist of a conversational context and counselor reflection. 350 of these observations constitute a reduced training dataset, randomly sampled from the full training set. 150 observations make up a randomly sampled test dataset. We ensured that the chat identifiers for messages in the test set uniquely differed from those included in the training set to avoid conversation overlap. Due to the sensitive nature of this dataset and privacy concerns, we cannot publicly share the client-counselor data, which has text from actual clients. A Hugging Face dataset card has been created and its contents have been included in the Appendix.

3.2 Model overview

To build an automatic reflection generation system, we use the Generative Pretrained Transformer 3 (GPT-3) architecture (Brown et al., 2020). With 175 billion parameters, GPT-3 is a language model able to mimic human text and is useful for dialogue generation (Zhang et al., 2020). We explore fine-tuning and few-shot learning approaches. Fine-tuning involves updating weights of a pre-trained model by training on a task-specific supervised dataset. Few-shot learning refers to the setting where the model is provided a task description and a few examples at inference time as conditioning, but weight updates are not allowed (Brown et al., 2020; Radford et al., 2019). We consider the few-shot approach, as scaling up language models greatly improves task-agnostic, few-shot performance (Brown et al., 2020). With the few-shot approach, there is a major reduction in the need for task-specific data and reduced concern of learning an overly narrow distribution, but it involves rapid adaptation to a new task with limited priming (Brown et al., 2020). Prompt-based few-shot learning on large language models achieves comparable results to state-of-the-art full-shot models in a variety of language understanding tasks, including for response generation using an empathetic dialogues dataset (Brown et al., 2020; Madotto et al., 2021).

Each example consists of the prompt, which is the conversational context that immediately precedes the counselor reflection. That is, it includes previous utterances from either the client or counselor up until and including the most recent prior client message that immediately followed a counselor's message. This ensures the client's expression following the previous counselor message(s) is included in the context. Given that reflection statements are often based on the most recent client message, and client messages on inspection of the data were often short, using this structure seemed appropriate. An instance illustrating the formatting of examples is provided in the Appendix. All examples adhere to the same format across the fine-tuned models and few-shot learning model.

We develop four reflection generation models:

Few-shot learning prompt-designed model. We design a prompt consisting of an instructional statement and 17 examples, each consisting of conversational context and a created reference reflection.

Fine-tuned model on reduced set of actual counselor responses. We fine-tune a model on 350 context-reflection examples where reflections are those of counselors of an online counseling service.

Fine-tuned model on comparable number of reference reflections. We fine-tune a model on 350 context-reflection examples where reflections are reference reflections created by one of the authors.

Fine-tuned model on full set of actual counselor responses. We fine-tune a model on 911 context-reflection examples where reflections are those of counselors of an online counseling service.

3.3 Creating reference reflections

The author who created 350 reference reflections for training and 150 for testing does not have a clinical psychology or medical doctorate but has undergone extensive training at mental health organizations including a crisis hotline service and a textline platform, totaling over 100 hours, which included one-on-one interactions and feedback from trained supervisors. In contrast, volunteers of the platform (from which the counseling data derives) receive online training that takes 45 minutes to 1 hour. Although the average word length was greater for reference reflections than actual counselor reflections, this was not intentional but likely a product of deliberate focus on communicating a reflection. Reflections were posed as questions in cases of limited context (e.g., "Ok, so it's fine. How are things feeling?") or uncertainty about the client's meaning (e.g., "It feels like everyone is disappointing you, is that right?"). Although the author attempted to vary responses, responses more frequently began with "It sounds like...", "I see, so...", "It seems...". In considering the context, client messages were given most weight in crafting reflections.

3.4 Prompt design

Based on experimentation within the OpenAI web interface to define a prompt structure that generated the most reasonable counselor reflections, we included a scenario description, i.e., a description of the nature of the requested response, and delimiters for the client and counselor. These decisions were influenced by previous work on prompt design. The Madotto et al. (2021) system for empathetic dialogues uses textual delimiters to distinguish interlocutors. Zheng and Huang (2021) found distinguishing input constructs (e.g., "User:" and "System:") is effective in boosting few-shot learning performance for grounded dialog generation tasks. For the discrete prompts, Zheng and Huang (2021) prepended input sequences with task instructions and found that discrete prompts generally outperform continuous prompts under few-shot and full data settings. We used a textual scenario description to guide the model to complete the reflection task:

> The counselor is a chatbot that listens empathetically, is kind, and reflects back how the user is feeling. The counselor reframes the client's message.

In preliminary experiments, we found providing the API with such instructions appeared to generate responses that more appropriately resembled reflections, as some responses when instructions were not included were more opinionated or brought the counselor's own struggles into conversation. We explored altering the instructions, considering e.g., "paraphrase", "reflect", "rephrase", but responses were similar, if not better, with "reframe". We experimented with the following variant: "The counselor is a chatbot that is empathetic, caring, and actively listens. The counselor reflects back the client's feelings and may offer direction." However, responses were similar but sometimes less relevant.

We used plain language to describe inputs and outputs, i.e., 'Counselor' and 'Client'. We tried other output descriptions, e.g., 'Therapist', but the results were very similar. We append '\nCounselor:' to the prompt to immediately precede the counselor completion. We explored other options that defined this response uniquely apart from any counselor messages in the context, e.g., 'Counselor reflection:', but performance did not appear to improve. For the token at the end of completions, we also tried using tokens that were more distinguishable from the content of examples, e.g.,'\n\n###\n\n', but '\n\n' appeared to exhibit better performance.

We explored different structures including continuation with and without a scenario description and continuation vs. a question-answer style. For continuation, the model would continue the conversation by completing the next system response. For question-answer style, the model is queried for what the system probably says next given the scenario description, and then the model answers the query with its predicted system response.

The question-answer style for this context had the following structure:

> Client: [message]
> To empathetically rephrase the client's message, what does the counselor probably say in response?
> Counselor:

We observed generated responses for a few held-out prompts and other client responses created by one of the authors and found the continuation with scenario description produced the best responses. With scenario description, the model generates more reflective and empathetic statements. While the question-answer style developed longer responses, responses were less relevant than reflections generated with the continuation structure.

In the prompt, we included 17 examples, i.e., the maximum number of held-out examples (distinct from the training set for fine-tuned models and the test set) that fit into the model's maximum context length with still enough tokens remaining to append the longest conversational context to be tested.

3.5 Hyperparameter selection and fine-tuning

We heuristically tuned the temperature and frequency penalty parameters. Temperature controls randomness; the frequency penalty controls how much to penalize new tokens based on their existing frequency so far. Reasonable values for the penalty coefficients are around 0.1 to 1 if the goal is to reduce repetitive samples somewhat without noticeably degrading sample quality. After exploring different levels, we found a temperature of 0.8 and frequency penalty of 0.8 were appropriate. A relatively high frequency penalty tended to lead to more complex reflections. The higher temperature also brought about more response diversity. The presence penalty controls the model's likelihood to talk about new topics. We set the presence penalty to 0. These selections are used across all models.

Models are created with the Davinci engine, as it is the most capable GPT-3 model and can perform tasks other models can, often with less instruction. text-davinci-001 is used for few-shot learning, and base model is used for fine-tuning. In preliminary experiments, we explored training with different epochs, including 2, 4, 7, 10. We found 4 epochs yielded the most optimal results on fifteen prompts not included in the test set, as the responses were more reflective, less directive, and less likely to infer context that could be in error. We used this selection across all fine-tuned models for comparable results.

3.6 Comparative experiments

We examine the impact of the following on generated reflection quality: few-shot learning versus fine-tuning, fine-tuning set size, and the quality of the fine-tuning set. We test overall preference for generated versus human responses and explore whether fine-tuning is associated with fluency degradation relative to few-shot learning.

We thus tested the following hypotheses, which were preregistered through the Center for Open Science prior to examining human evaluation data (O'Neil and Ungar, 2022).

Learning approach. The fine-tuned model on reference reflections will produce responses that have higher reflection resemblance ratings than the few-shot learning model (one-sided paired t-test).

Quantity for fine-tuning. The model fine-tuned on a larger set of actual counselor reflections from a counseling service will produce responses with higher reflection resemblance ratings than the model fine-tuned on the smaller set of actual counselor reflections (one-sided paired t-test).

Quality for fine-tuning. The model fine-tuned on reference reflections will produce responses with higher reflection resemblance ratings than the model fine-tuned on an equal number of actual counselor reflections (one-sided paired t-test).

Preference between human and computer. There will be no difference between overall preference for responses generated by the model fine-tuned on reference reflections (computer-generated) and reference reflections (human-generated) (two-sided paired t-test).

Fine-tuning degradation of fluency. There will be no difference between overall fluency for the fine-tuned model on reference reflections and the few-shot learning model (two-sided paired t-test).

3.7 Human evaluation

We recruited three annotators who have worked with this counseling service dataset with IRB approval. Although the evaluators are not clinically trained, they are highly familiar with the Motivational Interviewing Treatment Integrity and Motivational Interviewing Skill Code manuals and have experience labeling motivational interviewing counselor utterances for behavior codes including reflection. Although the author who created the reference reflections was involved in the model training process, the evaluators did not include this author and thus independently evaluated responses.

We administered a survey through Amazon Mechanical Turk Developer Sandbox. Each annotator evaluated outputs of the four models, the actual counselor reflection, and the reference reflection for 50 samples, a random subset of our test set. Provided with the conversational context, annotators evaluated the six responses based on fluency, resemblance of reflection, and overall preference.

Fluency refers to the response's overall fluency and human-likeness. The instructions noted non-capitalized words and colloquial language are acceptable and not to be considered fluency errors.

Reflection resemblance refers to whether the response captures and returns to the client something the client has said. *Overall preference* refers to the extent to which the evaluator likes the response.

We use a variation of Efficient Annotation of Scalar Labels (EASL), a hybrid approach between direct assessment and online pairwise ranking aggregation and rank-based magnitude estimation (Sakaguchi and Van Durme, 2018). Evaluators see all six responses at once (without knowledge of their origin) and use a 1 to 5 sliding scale to rate responses on each dimension. The order of model responses for each context was randomized. We provided example response ratings for ratings of 1 and 5 on overall fluency and reflection resemblance but not overall preference, noting its subjectivity.

Evaluation of overall preference and consideration of humanness in measuring fluency were influenced by Smith et al. (2022), which adapted metrics from Li et al. (2019). The reflection resemblance and fluency criteria are loosely similar to that of Shen et al. (2020). Reflection resemblance slightly differs from their description of reflection-likeness, as we do not explicitly reference paraphrasing or summarizing; also, our notion of fluency highlights the extent to which responses are human-like. Fluency was also evaluated for generated empathetic responses by Majumder et al. (2020).

4 Results

4.1 Human evaluation

The average rating for each response source on overall fluency, reflection resemblance, and overall preference are shown in Figure 2. Average ratings for the counselor responses are low in relation to the reference reflections. Moreover, it appears to be better to provide relatively higher quality examples for fine-tuning, as seen by the superior performance on all criteria of the few-shot learning model and the model fine-tuned on reference reflections compared to the models fine-tuned on counselor examples.

Using Krippendorff's alpha (ordinal method), we measured inter-annotator agreement (Krippendorff, 2018). We obtained alpha values of -0.0369, 0.557, and 0.358 for overall fluency, reflection resemblance, and overall preference, respectively. Although these agreement values are low, 0.557 for reflection resemblance is notably higher than the 0.23 agreement for reflection-likeness in the most relevant prior work Shen et al. (2020).

There are a few considerations for the low agreement. As fluency ratings are high across all models, the chance correction agreement is low. Potential contributions to the low agreement include the subjectivity of "human-like" and the measure's lack of specificity. Disagreement appeared to arise in the presence of colloquial language or minor misspellings or missing apostrophes in contractions. It is possible annotators incorporated more subjective quality assessments to varying extents given that the criterion in part was evaluating human-likeness. The subjective nature of overall preference is likely the primary reason for the fair agreement for this criterion. Differences in preference and varying knowledge on reflections may have contributed to the moderate agreement on reflection resemblance.

The Pearson correlations between criteria are as follows: 0.367 for fluency and reflection resemblance, 0.341 for fluency and preference, and 0.699 for reflection resemblance and preference. We would expect a reasonable correlation between reflection resemblance and preference, as a more reflective statement is likely to be more appealing and feel more meaningful. Given the correlated criteria, we measured inter-annotator agreement for annotators' average rating (i.e., for each annotator, we averaged that annotator's ratings for each model response), reflecting an overall quality measure for each annotator. The inter-annotator agreement alpha for their average judgments is 0.505.

We conducted paired t-tests as specified in Section 3.6. Two tests are associated with significant results. Fine-tuning on higher quality examples produces responses that better resemble reflections (p-value < 2.2e-16), and there is a difference between the overall preferability of computer-generated responses and human responses (p-value = 1.78e-05).

4.2 Qualitative examples

To illustrate conversational contexts and associated reflections based on the six sources, we present three representative qualitative examples in Table 1. The first example was selected to highlight the more nuanced inferior quality of a reflection produced by the model fine-tuned on 911 counselor responses. The second example illustrates clear faults of a reflection produced by the model fine-tuned on 350 counselor responses. The third example features a natural limitation of lengthier responses.

The models fine-tuned on 911 counselor responses and 350 counselor responses tend to pro-

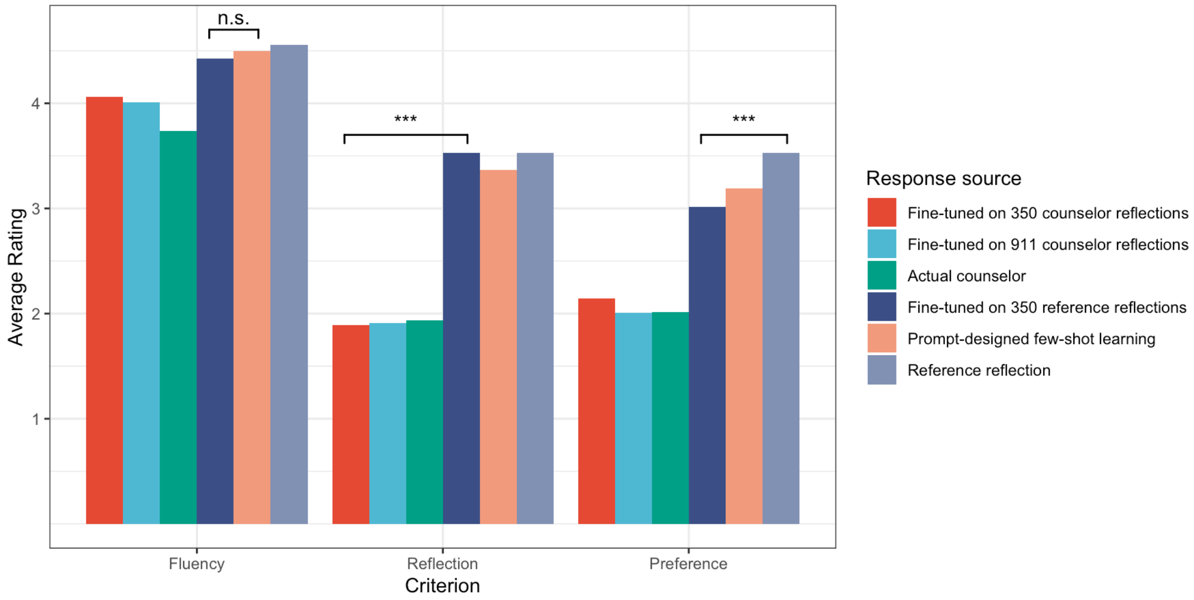


Figure 2: Human evaluation average ratings on overall fluency, reflection resemblance, and overall preference. Standard errors on estimates range from 0.057 to 0.111. Three paired t-tests results are noted. Although not defined in our comparative experiments, there is no statistically significant difference between reflection resemblance ratings of the few-shot learning model responses and reference reflections but a significant difference in overall preference.

duce less accurately reflective and less substantive responses (with shorter average reflection length) compared to the model fine-tuned on 350 reference reflections and the few-shot learning model.

For the first conversational context in Table 1, the client is expressing the subject's lack of consideration rather than a statement about differently weighing the feelings of the subject and client, as suggested by the response of the fine-tuned model on 911 counselor responses. The reflections of the model fine-tuned on 350 reference reflections and the few-shot learning model are more accurate. For the second context, the response of the fine-tuned model on 350 counselor responses fails to reflect back the heavy emotion expressed. In contrast, the reference reflection and the responses of the fine-tuned model on 350 reference reflections and the prompt-designed few-shot learning model capture the client's statement and more sensitively address the client's distress than the actual counselor, supporting the importance of tuning or conditioning on high-quality responses. Moreover, for the third context, the response of the model fine-tuned on 911 counselor responses and the actual counselor response are not very substantive. In contrast, the responses of the model fine-tuned on 350 reference reflections and the few-shot learning model capture that the client's pain has sustained over time. The response of the model fine-tuned on 350 reference reflections addresses the medical issues raised but infers these are able to be, and have not yet been, "figured out", which is not necessarily true. Evidently, with lengthier responses, there is greater potential additional assumed context is invoked.

5 Discussion

We introduced a system for generating reflections in peer-to-peer counseling and explored the impacts on generated output quality of few-shot learning and fine-tuning, fine-tuning on a larger training set, and fine-tuning on higher-quality instances. Our human evaluation shows humans rate responses generated by a model fine-tuned on higher-quality examples as more reflection-like, and there is a significant difference between overall preference for human and computer-generated responses. There is no evidence to suggest fine-tuning GPT-3 results in better reflections than few-shot learning. Our results highlight the care needed in selecting examples for fine-tuning a model to generate reflections. Evidently, input that reveals the nature of the counselor-client dialogue context alone is not enough to produce high-quality model responses, but also reflections epitomizing effective responses. Our results also suggest generative models tuned on

Table 1: Qualitative examples of conversational contexts and the associated reflection responses.

Conversational context	Response source	Response
Client: I'm 19, he's 25. He's not very considerate of how I feel but says he cares about me and loves me. Counselor:	Online counseling service actual counselor	The words are easy, actions are needed. Guys who are 25 just desire to have different experiences.
	Reference reflection	It sounds like his words suggest that he cares but his actions indicate otherwise.
	Fine-tuned on 350 actual counselor responses	So he's 25, but not very considerate of your feelings, and you say he's not very mature
	Fine-tuned on 911 actual counselor responses	Well in a relationship like that, you are going to be more concerned about how he feels than how you feel
	Fine-tuned on 350 reference reflections	It sounds like he claims to care but you feel he is not very thoughtful.
	Prompt-designed few-shot learning	It sounds like he does not always take your feelings into account, but loves you.
Client: Things have been stressful. I have been going through a hard time and my anxiety has reached the ceiling hahah Client: but ok haha Counselor:	Online counseling service actual counselor	It doesn't sound like you're okay
	Reference reflection	It doesn't sound like you're fine. It is okay to be having a hard time dealing with a lot of stress. I'm glad you reached out.
	Fine-tuned on 350 actual counselor responses	Ooh I see, well it seems you've gotten through it!
	Fine-tuned on 911 actual counselor responses	Yeah definitely been rough
	Fine-tuned on 350 reference reflections	It sounds like you have been going through a lot and having a hard time.
	Prompt-designed few-shot learning	It sounds like you have been going through a lot of stress lately. It is understandable that your anxiety has increased.
Client: Not much, a little maybe Client: But I have many genetic medical problems Client: I've had pain all my life Counselor:	Online counseling service actual counselor	That's just labels
	Reference reflection	I can't imagine how challenging it must be to deal with pain for so long, having several medical challenges as you mention.
	Fine-tuned on 350 actual counselor responses	I understand. It's very brave of you to continue living in the way you have been, despite all these issues.
	Fine-tuned on 911 actual counselor responses	that must be hard
	Fine-tuned on 350 reference reflections	I am hearing that you have been in pain for a long time, and it sounds like there are medical issues that have not been figured out yet.
	Prompt-designed few-shot learning	It sounds like you have been through a lot in your life.

high-quality reflections generate better reflections than the actual peer counselors. However, with respect to the difference in quality between actual counselor and reference reflections, it is important to recognize that the author knew the reference reflections were being evaluated given minimal context, whereas naturally, the actual counselors may not have such a focus on creating a well-crafted response for any given moment in conversation.

The global burden of mental illness is significant (Collins et al., 2011). Online mental health support platforms with peer counselors are a means of scaling up support, but the challenge of effective training remains. Our work represents how natural language generation can be used to help support peer counselors. Such a system with a machine-in-the-loop approach can provide actionable suggestions to counselors and in turn potentially offer those seeking help more reflective support.

Our results also have potential implications for the wider context of online peer interactions, e.g., peer grading and customer support. Many contexts require responses to be consistent with a particular style; our results indicate higher-quality examples are critical for fine-tuning. Given we did not find evidence to suggest fine-tuning GPT-3 results in better quality reflections than few-shot learning, we suggest future work further explore the trade-off between quality and quantity of examples provided to orient models toward the domain of interest.

Future work could build a collaborative writing tool, e.g., Clark and Smith (2021), for reflections and study the extent to which counselors accept, modify, or abandon suggestions. Future work could explore inclusion of an additional input for the counselor to provide conversational redirection. Our approach could also be applied to build and evaluate generative systems for other clinical skills.

6 Limitations

As a result of our decision to limit conversational context to most recent messages, sometimes actual counselor responses took into account more of the conversation than was captured in the prompt. In Shen et al. (2020), the context window size was five utterances, and a larger window size did not improve performance in preliminary experiments. However, it may be worth further exploring how greater context could enable more complex reflection statements. Current language models lack the ability to account for the broader context. Another technical limitation of this work entails the use of only one type of large language model for the reflection task. The experiments in this work are targeted at comparing few-shot learning and fine-tuning as well as assessing the impact of the quality of examples provided for tuning, and so in the interest of narrowing the focus, this work lacks a comparison of the quality of responses using various large language models in the counseling setting.

The evaluation criteria used in the human evaluation have their own limitations in that these dimensions do not necessarily reflect what is most therapeutically beneficial in the counseling setting and what offers the best experience for clients. The models were evaluated on their ability to generate fluent reflections and not on true therapeutic impact. Moreover, given the low agreement among annotators, the criteria's limited specificity likely introduced ambiguity and different interpretations in rating responses. Future work should consider having annotators go through a first round of annotation followed by discussion of disagreements with opportunity to clarify judgments and resolve different interpretations of the criteria, thus offering a means of potentially reducing disagreement for the subsequent annotation process.

The reference reflections in this work were created by one of the authors, whose experience with counseling and motivational interviewing derives from over one hundred hours of training at a crisis hotline and textline service and experience through a fellowship developing and user testing a platform for nurses to practice and grow their motivational interviewing skills. Therefore, these reflections may not be as clinically precise as are possible from a medical professional, and the diversity of reflections is inherently limited. Additionally, this work examined one mental health support community; peer supporters of this counseling service receive more training than some online support groups, where members do not receive training, but substantially less training than suicide hotline volunteers may receive.

7 Ethics

GPT-3 was trained on over 45 terabytes of data from the internet and books, and large volumes of data collected from online sources will inevitably contain biases. There may thus be inadvertent discrimination against subclasses of particular protected groups. Using generated responses as a source of guidance rather than using generative systems as the counselors themselves may be able to help balance the benefits and risks of using artificial intelligence in delicate mental health settings. It is critical such systems are not misused by companies seeking to maximize efficiency and minimize cost.

Such a tool cannot replace counselor training, as it remains critical for counselors to be able to adequately assess responses prior to using them, particularly so that if generated text is biased or careless, it is reviewed and discarded. Thus, it is necessary counselors continue to receive sufficient training to ensure they can identify clearly inappropriate generated text. When such technology is introduced to counselors, its limitations should be clearly communicated and its use monitored. Additionally, it is imperative deployment of and subsequent experimentation with such a tool is done only with informed consent of users of an online counseling service. Importantly, we see such automated tools as a way of assisting online counselors, especially peer counselors, not as replacing humans.

Acknowledgements

We are grateful to the annotators for their participation in the human evaluation. We also thank Robert Kraut for his input and guidance in the framing of this research and the analysis.

References

David C. Atkins, Mark Steyvers, Zac E. Imel, and Padhraic Smyth. 2014. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):1–11.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611, Florence, Italy. Association for Computational Linguistics.

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 *23rd International Conference on Intelligent User Interfaces*, IUI '18, page 329–340, New York, NY, USA. Association for Computing Machinery.

Elizabeth Clark and Noah A. Smith. 2021. Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3566–3575, Online. Association for Computational Linguistics.

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*, 475(7354):27–30.

Antonia S. Csillik. 2013. Understanding motivational interviewing effectiveness: Contributions from rogers' client-centered approach. *The Humanistic Psychologist*, 41(4):350–363.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 129–133, Prague, Czech Republic. Association for Computational Linguistics.

Sangdo Han, Kyusong Lee, Donghyeon Lee, and Gary Geunbae Lee. 2013. Counseling dialog system with 5W1H extraction. In *Proceedings of the SIGDIAL 2013 Conference*, pages 349–353, Metz, France. Association for Computational Linguistics.

Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982*.

Zac E. Imel, Mark Steyvers, and David C. Atkins. 2015. Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy*, 52(1):19–30.

Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, Kang Min Yoo, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. *arXiv preprint arXiv:2205.12685*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*, 4th. edition. Sage Publications, US.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Michael B. Madson, Andrew C. Loignon, and Claire Lane. 2009. Training in motivational interviewing: A systematic review. *Journal of substance abuse treatment*, 36(1):101–109.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

William R. Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*, 3rd. edition. Guilford Press, New York, NY.

William R. Miller, Carolina E. Yahne, Theresa B. Moyers, James Martinez, and Matthew Pirritano. 2004. A randomized trial of methods to help clinicians learn motivational interviewing. *Journal of Consulting and Clinical Psychology*, 72(6):1050–1062.

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*, 10:746.

Chiaki Miyazaki. 2023. Dialogue generation conditional on predefined stories: Preliminary results. *IEEE Access*.

Theresa B. Moyers, Tim Martin, Jon M. Houck, Paulette J. Christopher, and J. Scott Tonigan. 2009. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *Journal of Consulting and Clinical Psychology*, 77(6):1113–1124.

Mark Olfson. 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Affairs*, 35(6):983–990.

Emma O'Neil and Lyle H Ungar. 2022. Reflection generation for peer-to-peer counseling setting using generative pre-trained transformer architecture.

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

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics.

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

Siqi Shen, Veronica Perez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3096–3107, Dublin, Ireland. Association for Computational Linguistics.

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

Eric Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. 2022. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 77–97, Dublin, Ireland. Association for Computational Linguistics.

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

Anuradha Welivita and Pearl Pu. 2023. Boosting distress support dialogue responses with motivational interviewing strategy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5411–5432, Toronto, Canada. Association for Computational Linguistics.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Chujie Zheng and Minlie Huang. 2021. Exploring prompt-based few-shot learning for grounded dialog generation. *arXiv preprint arXiv:2109.06513*.

A Format of examples

Below is an instance where the prompt contains a single client message. Note that "prompt" and "completion" are the names of the fields requiring designation for fine-tuning:

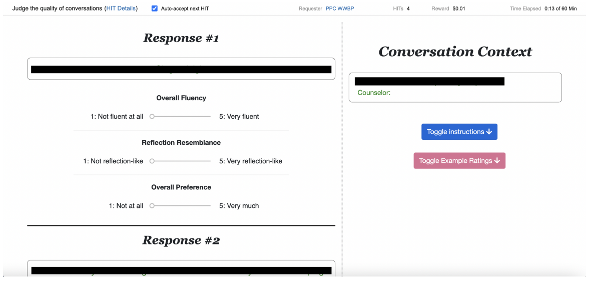
`{"prompt":"Client: <message>\nCounselor:", "completion":" <reflection>\n\n"}`

There can alternatively be multiple client messages (and also counselor messages) before the counselor reflection:

{"prompt": "Client: <message$_1$>\nClient: <message$_2$>\nCounselor:", "completion":" <reflection>\n\n"}

B Survey screenshots

Figure 3 illustrates the user interface for the survey annotators completed for the human evaluation, and Figure 4 illustrates the instructions users could toggle throughout the survey. Users could also toggle example ratings, but these examples have been omitted given that the text of the examples themselves would need to be redacted.



Judge the quality of conversations (HIT Details) ☑ Auto-accept next HIT Requester PPC WWBP HITs 4 Reward $0.01 Time Elapsed 0:13 of 60 Min

Response #1

[redacted]

Overall Fluency

1: Not fluent at all ○ 5: Very fluent

Reflection Resemblance

1: Not reflection-like ○ 5: Very reflection-like

Overall Preference

1: Not at all ○ 5: Very much

Response #2

[redacted]

Conversation Context

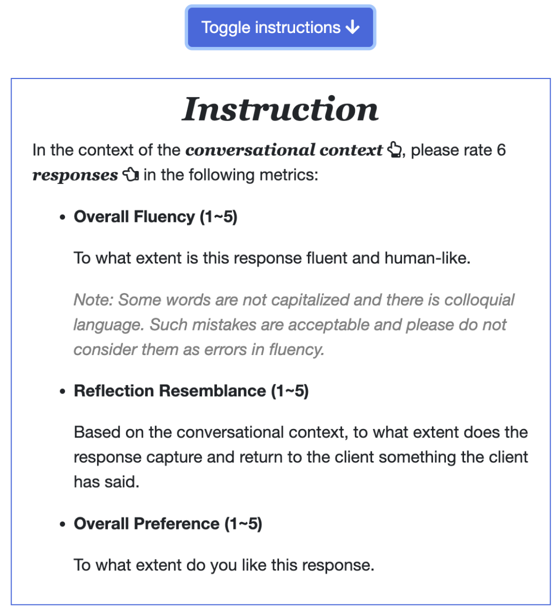
[redacted]

Counselor:

Figure 3: Mechanical Turk Developer Sandbox survey user interface with the text redacted given privacy limitations. Note that all six responses for a given conversational context were presented on the screen in a scrollable view.

C Data anonymity

Chat identifiers for conversations were only used to avoid overlap of conversations between the training and testing sets. Only the messages themselves and not the chat identifiers were used for fine-tuning and prompting GPT-3. The conversational contexts and online counseling service actual counselor responses in Table 1 of the paper have been altered due to privacy limitations so they are not the identical messages themselves. The messages were put



Instruction

In the context of the ***conversational context*** 👆, please rate 6 ***responses*** 👇 in the following metrics:

- **Overall Fluency (1~5)**

  To what extent is this response fluent and human-like.

  Note: Some words are not capitalized and there is colloquial language. Such mistakes are acceptable and please do not consider them as errors in fluency.

- **Reflection Resemblance (1~5)**

  Based on the conversational context, to what extent does the response capture and return to the client something the client has said.

- **Overall Preference (1~5)**

  To what extent do you like this response.

Figure 4: The toggled instructions annotators had access to as they evaluated the responses. Also provided was an example of a conversational context and examples of response ratings for overall fluency and reflection resemblance given the conversational context.

through a full round of Google Translate and additionally modified by one of the authors, while being careful not to introduce different semantics, to ensure the presented messages appear sufficiently different from the originals.

D Hugging Face Dataset Card

D.0.1 Dataset Summary

The dataset derives from conversations between clients and counselors on a large peer-to-peer online counseling service. There are a total of 1061 observations across training and testing datasets, with 50 additional randomly sampled examples used in defining the few-shot learning prompt or for validation purposes in tuning hyperparameters, thus totaling 1111 observations across these sets. These observations were sourced from a larger dataset consisting of annotations of several clinical counseling skills. Messages were annotated at utterance level with counselor verbal behaviors using the Motivational Interviewing Treatment Integrity 4.2 (MITI) and the Motivational Interviewing Skill Code 2.5 (MISC) manuals. Our focus is on counselor reflections. Thus, the dataset consists of conversational context-counselor reflection pairs.

D.0.2 Supported Tasks and Leaderboards

The dataset was used for conditioning and tuning generative models for generating reflection statements in the domain of peer-to-peer counseling.

D.0.3 Languages

The language in the dataset is English.

D.1 Dataset Structure

D.1.1 Data Instances

Each instance consists of the chat room id of the conversation in which the dialogue occurred, the prompt which is the conversational context that immediately precedes the counselor reflection (including previous utterances from either the client or counselor up until and including the most recent prior client message that immediately followed a counselor's message), and the completion, which is the counselor reflection.

```
{
  'chat_id': "1234567",
  'prompt': "Client: I'm 19, he's 25.
           He's not very considerate
           of how I feel but says he
           cares about me and loves
           me.\nCounselor:",
  'completion': " The words are easy,
               actions are needed.
               Guys who are 25 just
               desire to have
               different
               experiences.\n\n",
}
```

D.1.2 Data Fields

chat_id: an integer defining the chat id of the conversation. *prompt*: a string corresponding to the conversational context preceding the counselor reflection with the messages separated by new line characters and each utterance prepended by 'Client:' or 'Counselor:'. The string ends with 'Counselor:' to indicate that it is followed by the counselor completion. *completion*: a string corresponding to the counselor reflection.

D.1.3 Data Splits

The dataset is split into training, testing, and a small set of 50 examples used either for designing the few-shot learning prompt or tuning hyperparameters. For prompt design, the structure of the prompt with examples was influenced by prior work. Thus, prior work provided scaffolding for our approach;

the selections made within these frameworks were driven by exploration. The hyperparameters were tuned heuristically given the essential qualitative nature of reflection evaluation in the counseling context. 911 examples were used for training. 350 of these examples also constitute a reduced training set used in comparative experiments. 150 examples were used for testing. 50 of these testing examples (randomly selected) were used in the human evaluation. We ensured that the chat identifiers for messages in the test set uniquely differed from those included in the training set.

D.2 Dataset Creation

D.2.1 Curation Rationale

Reflective listening is a critical skill in peer-to-peer counseling that is only effective when tailored to the context. Thus, we wanted to home in on this particular skill and explore the potential of state-of-the-art language models for text generation in this domain. GPT-3 was used in this work given the model was trained on a larger dataset and has many more parameters than other LLMs at the time of experimentation. Collaborative generation could be a key tool for online peer support. As per other work (Sharma et al., 2021), we think this will be an application area of great societal benefit.

D.2.2 Source Data: Initial Data Collection and Normalization

The dataset was created by filtering the larger dataset of utterances annotated for many different counseling skills to only those counselor messages annotated as reflections. Then, the prompt instances were created by identifying the preceding messages for each of these counselor reflection instances. After the prompts were initially created, prompts with less than or equal to five words were removed.

One of the authors created reference reflections for each of the 350 training example prompts in the reduced training set and each of the 150 testing example prompts. The reference reflections were created based on the author's experience in volunteering as a counselor at crisis hotlines.

D.2.3 Source Data: Who are the source language producers?

The 'client' messages are utterances of those seeking mental health support on a large online counseling service platform. The 'counselor' messages are utterances of minimally-trained peer counselors of this large online counseling service.

For each of the 350 training example prompts in the reduced training set and each of the 150 testing example prompts, a reference reflection was also created by one of the authors.

D.2.4 Annotations: Annotation process

The human evaluation examined text of generative models fine-tuned on the full training set, a reduced training set, and reference reflections; a few-shot learning model; the actual counselor; and the reference reflection.

We administered a survey through Amazon Mechanical Turk Developer Sandbox. 50 testing prompts, a random subset of our test set, were provided along with the corresponding six response sources. Provided with the conversational context, the annotators evaluated responses based on three criteria: fluency, resemblance of reflection, and overall preference. Thus, for each context, evaluators measured the fluency, reflection resemblance, and overall preference for all six candidate responses. The three criteria for evaluation were motivated by prior work.

We used a variation of Efficient Annotation of Scalar Labels (EASL), a hybrid approach between direct assessment and online pairwise ranking aggregation and rank-based magnitude estimation (Sakaguchi and Van Durme, 2018). Evaluators saw all six responses at once (without knowledge of each response's origin) and used a sliding scale from 1 to 5 to rate the responses based on each of the three dimensions. The order of the model responses for each conversational context was randomized. We provided examples of response ratings for ratings of 1 and 5 on the overall fluency and reflection resemblance dimensions. However, we did not include an example for overall preference, noting its subjectivity.

Fluency refers to the response's overall fluency and human-likeness. In the instructions, we noted non-capitalized words and colloquial language are acceptable and not to be considered fluency errors. Reflection resemblance refers to whether the response captures and returns to the client something the client has said. Overall preference refers to the extent to which the evaluator likes the response.

Using Krippendorff's alpha, we measured inter-annotator agreement, obtaining alpha values of -0.0369, 0.557, and 0.358 for overall fluency, reflection resemblance, and overall preference, respectively. Although these agreement values are low, the 0.557 inter-annotator agreement we obtained for reflection resemblance is notably higher than the inter-annotator agreement obtained for reflection-likeness in the most relevant prior work Shen et al. (2020).

D.2.5 Annotations: Who are the annotators?

The three annotators recruited for the human evaluation were familiar with counseling reflections. All three annotators have worked with this large online counseling service dataset with IRB approval. They are computer science students in the United States; two annotators are graduate students, and one annotator is an undergraduate student. Two annotators are female, one is male. The annotators are highly familiar with the Motivational Interviewing Treatment Integrity and Motivational Interviewing Skill Code manuals and have experience labeling MI counselor utterances for various behavior codes including reflection. They were compensated through payment. Each annotator received $25. They each took about two hours to complete all survey HITs, thus equating to roughly $12.50 per hour in compensation. Annotators were instructed that their ratings were part of a human evaluation study that entailed measuring the quality of automatically generated reflection responses and human-generated responses.

D.2.6 Personal and Sensitive Information

Due to the sensitive nature of this dataset and privacy concerns, we cannot share prompts (conversational contexts), which have text from actual clients (confidential). The dataset was shared by a counseling service for research purposes. The annotators recruited for the human evaluation have worked with this dataset with IRB approval.

D.3 Considerations for Using the Data

D.3.1 Social Impact of Dataset

This dataset of reflections in peer-to-peer counseling can be used as a reference point in understanding and evaluating counselor clinical skills and furthering the potential of language technology to be applied in this space. Given the sensitive nature of the mental health care context and the minimal training of these counselors, the use of such data requires care and understanding of the limitations of technology defined based on this language.

D.3.2 Discussion of Biases

Much of the language of conversations on this online counseling service platform is very informal, and some client and counselor utterances may also contain pejorative language.

As for the generated text assessed in the human evaluation of this work, it is important to note that GPT-3 was trained on over 45 terabytes of data from the internet and books, and large volumes of data collected from online sources will inevitably contain biases that may be captured. There may thus be inadvertent discrimination against subclasses of particular protected groups. Using generated responses as a source of guidance rather than using generative systems as the counselors themselves may be able to balance the benefits and risks of using artificial intelligence in delicate mental health settings. It is imperative that such systems are not misused by companies seeking to maximize efficiency and minimize cost.

The reference reflections in this work were created by one of the authors, whose experience with counseling and motivational interviewing derives from over one hundred hours of training at a crisis hotline and textline service and experience through a fellowship developing and user testing a platform for nurses to practice and grow their motivational interviewing skills. Therefore, the reference reflections may not be as clinically precise as are possible from a medical professional, and the diversity of reflections is inherently limited.