IMBUE: Improving Interpersonal Effectiveness through Simulation and Just-in-time Feedback with Human-Language Model Interaction

Inna Wanyin Lin[1] **Ashish Sharma**[1] **Christopher Michael Rytting**[1]
Adam S. Miner[2] **Jina Suh**[3] **Tim Althoff**[1]
[1]Paul G. Allen School of Computer Science & Engineering, University of Washington
[2]Stanford University [3]Microsoft Research
ilin@cs.washington.edu

Abstract

Navigating certain communication situations can be challenging due to individuals' lack of skills and the interference of strong emotions. However, effective learning opportunities are rarely accessible. In this work, we conduct a human-centered study that uses language models to *simulate* bespoke communication training and *provide just-in-time feedback* to support the practice and learning of interpersonal effectiveness skills. We apply the interpersonal effectiveness framework from Dialectical Behavioral Therapy (DBT), DEAR MAN, which focuses on both conversational and emotional skills. We present IMBUE, an interactive training system that provides feedback 25% more similar to experts' feedback, compared to that generated by GPT-4. IMBUE is the first to focus on communication skills and emotion management simultaneously, incorporate experts' domain knowledge in providing feedback, and be grounded in psychology theory. Through a randomized trial of 86 participants, we find that IMBUE's simulation-only variant significantly improves participants' self-efficacy (up to 17%) and reduces negative emotions (up to 25%). With IMBUE's additional just-in-time feedback, participants demonstrate 17% improvement in skill mastery, along with greater enhancements in self-efficacy (27% more) and reduction of negative emotions (16% more) compared to simulation-only. The improvement in skill mastery is the only measure that is transferred to new and more difficult situations; situation-specific training is necessary for improving self-efficacy and emotion reduction.

1 Introduction

Some conversations can be challenging to navigate (Stone et al., 2023), whether they concern negotiating a salary increase with a boss, discussing healthcare options with an aging parent, or asking a friend to return the money they owe. Various communication frameworks assist individuals in conducting such conversations by providing a set of skills to apply (Stone et al., 2023; Rosenberg and Chopra, 2015; Linehan, 2014; Hartley, 2002).

However, psychology research highlights that a lack of communication skills is not the only obstacle to effective communication, particularly in emotionally charged situations (Linehan, 2014). Difficult conversations can evoke strong emotions that disrupt effective communication, even for individuals with solid communication skills (Luff et al., 2016; Henderson, 2016). To successfully communicate during challenging situations, it is crucial to focus not only on communication skills but also on managing emotions.

The popular DEAR MAN framework, from Dialectical Behavioral Therapy (DBT), was originally developed for Borderline Personality Disorder, but is widely used to teach conversational strategies and emotional regulation (Linehan, 2014). It includes conversational strategies (*Describe, Express, Assert, Reinforce,* and *Negotiate*) and a desired "state of mind" (*Mindful* and *Confident*) for productive conversations. Remaining mindful and confident in challenging conversations helps speakers regulate difficult emotions so they can successfully exercise their conversational strategies (§2;A).

Currently, DEAR MAN skills are mainly taught in therapy sessions and practiced either onsite in a roleplaying setting or at home with paper worksheets, which presents several challenges. Access to a trained therapist may be limited due to the significant shortage of mental health professionals (Olfson, 2016). Outside of therapy sessions, static worksheets do not provide opportunities for interactive role-playing and just-in-time feedback necessary for effective learning (Beck, 1979; Gagne, 1965; Beck, 1996).

Prior work in NLP has shown the ability of LMs to simulate personas and social interactions (Argyle et al., 2023b; Park et al., 2022, 2023). A few recent works leverage this capability by using LMs to

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 810–840
August 11-16, 2024

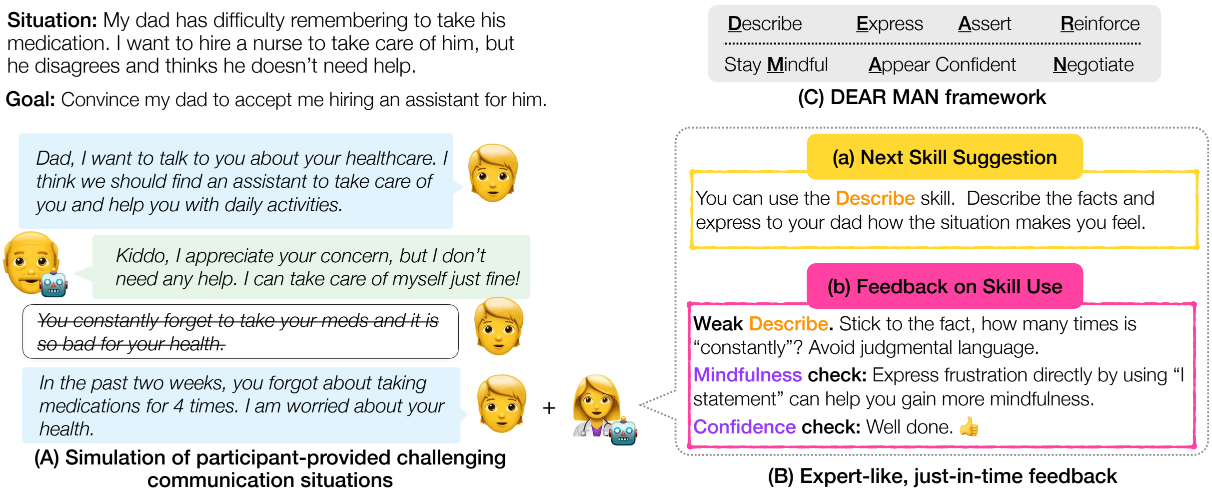


Figure 1: Overview of IMBUE, an interactive training system that **(A)** simulates bespoke communication situations and **(B)** provides expert-like just-in-time feedback based on **(C)** the DEAR MAN framework. IMBUE is backed by LMs that perform two tasks: **(a)** Next skill suggestion: before a user writes a message, IMBUE suggests a skill to apply (§4.2). **(b)** Feedback on skill use: after a user writes a message, IMBUE provides skill rating and improvement suggestions (§4.1).

help people improve interpersonal skills (Liu et al., 2023b) or conflict resolution skills (Shaikh et al., 2023), without considering emotional regulation.

Our work extends this literature, focusing on communication and emotional regulation skills simultaneously, incorporating expert domain knowledge into feedback, and grounding strategies in clinical psychology theory. We conduct a human-centered study and make three key contributions.

First, we present a formative study and an expert annotated dataset on DEAR MAN skill use. We conduct a formative study to gain insights from psychology experts on best practices when simulating challenging conversations and providing fine-grained feedback (§2). To understand how clinicians provide feedback on DEAR MAN in their practice and to develop and evaluate our method on real situations, we collect a dataset from crowd workers consisting of difficult situations they encounter and simulated conversations within them (the crowd worker being paired with a role-playing LM partner). We then ask psychology experts specifically trained in teaching DBT skills to annotate these conversations, assessing skill use and offering suggestions for improvement (§3).

Second, we develop computational methods to provide feedback using insights from the formative study and collected dataset (§4). We propose a new prompting strategy, demonstrating contrasting pairs of strong and weak utterances, in addition to state-of-the-art prompting methods. Our method improves the accuracy in skill use evaluation, outperforming GPT-4 by 24.8%, and more expert-like, specific, and actionable improvement suggestions.

Third, we build IMBUE, an interactive training system that *simulates* difficult conversations and *provide just-in-time feedback* backed by LMs to support the practice and learning of DEAR MAN skills (Figure 1). IMBUE can be used at an individual's convenience to practice both communication and emotional regulation. Through a randomized trial with 86 participants, we evaluate IMBUE 's training outcomes on skill mastery, emotion reduction, and self-efficacy (§6). We show that a simulation-only variant of IMBUE improves participants' self-efficacy towards having the conversation–boosting confidence (27%) and reducing worries (4%)–and emotion reduction towards the situations–reducing fear (16%) and sadness (12%)–while not improving skill mastery significantly. With the addition of just-in-time feedback, the participants' skill mastery significantly improved by 17.6%, with additional improvement in self-efficacy (confidence, 26.7%) and emotion reduction (fear, 15.7%).

2 Formative Study to Inform Design

To understand how DEAR MAN skills are taught in practice, we conduct a formative study with three clinical experts, summarizing crucial insights and corresponding design decisions below. Further details on the study procedure are in Appendix B.

Insight 1: Guide clients to focus on facts instead of making judgmental comments when describing a situation. We refrain from asking the participants to describe the personality of the conversation partner, even though it may help LM simulate a more realistic conversation, and instead only focus on past *behaviors* that might influence the difficulty of the situation in IMBUE.

Insight 2: Among the DEAR MAN skills, D, E, A, R, N are conversation strategies one can choose for each utterance. Mindful and Confident are the "state-of-mind". One should always stay mindful and confident throughout the entire conversation. Therefore, in each turn, IMBUE gives participants the option to choose from five conversation strategies. Conversation skills are evaluated only if they are selected for use, while mindfulness and confidence are assessed for each utterance (§5;§6).

Insight 3: Practicing simpler or less emotionally intense situations helps with harder situations. We collect difficulty levels from participants and use the less difficult situation in training (§6).

Insight 4: Training should prioritize emotion management. It is not considered a successful use of DEAR MAN skills if the client negotiates well but gets agitated. We evaluate training outcomes in three aspects: skill mastery, emotion reduction, and self-efficacy (§6).

Insight 5: Choosing strategies before writing helps learning. We adapt this design in data collection (§3) and IMBUE.

3 Data Collection

We collect a dataset to understand how clinicians provide feedback on DEAR MAN in their practice and to develop and evaluate our method.

Data collection with Crowdworkers. We intend for our system to be used by individuals without specialized knowledge, so we collect data from crowdsourcing platforms. We recruit 20 people from Amazon mTurk who provided 60 different situations and annotations. Each worker provides three conversations, one from each of the following categories: family, social, and work. Workers are asked to have conversations with our LM, which was instructed to roleplay as their partner during these conversations. Each conversation needs to be at least 10 responses from the worker, or until the simulated conversation partner "agrees" with the worker, whichever comes first. In each utterance, workers need to select one or more strategies they want to use in the given utterances to encourage them to follow the DEAR MAN framework as much as they can (§2). We include more ethics and safety details in §8;I.

DEAR MAN expert annotation. Following a similar recruitment process for the formative study, we recruited another six clinical experts who have received specialized training and actively practiced DBT. We only select those who indicate they "sometimes" or "regularly" work with clients on DEAR MAN skills on the signup form §C. Each expert annotated 2 to 4 conversations randomly selected from the dataset. In the final dataset, we have 18 conversations annotated, containing 163 utterances in total. For each utterance in a conversation, experts provided annotations on 1) select the skills identified in the utterance, 2) rate the skill use with one of `strong` or `weak`[1], 3) for `weak` ratings, indicate suggestion for improvement and provide a rewritten utterance, 4) for skills not used, indicate reasons to use if the expert suggests to use them and provide a rewritten utterance. The interface for this annotation is shown in Appendix L.

4 Methodology

IMBUE is an interactive training system that simulates bespoke communication situations and provides expert-like just-in-time feedback based on the DEAR MAN framework. IMBUE is backed by LMs that perform two tasks: (a) Next skill suggestion: before a user writes a message, IMBUE suggests a skill to apply (§4.2), (b) Feedback on skill use: after a user writes a message, IMBUE provides skill rating and improvement suggestions (§4.1). We describe our methods for performing these tasks.

4.1 Skill rating and improvements suggestions

To ensure low latency and cost-efficiency, we define a multitask problem: for a situation S, an utterance U_i, and a skill L_i, simultaneously generate skill rating R_i and improvement suggestions F_i. The major challenges include operationalizing complex DEAR MAN constructs grounded in psychology and supporting the variety of situations users may want to simulate. Previous research has shown the effectiveness of in-context learning for various NLP tasks (Brown et al., 2020; Sharma et al., 2023b). Our method builds on these approaches with four key components: 1) curated rubrics to augment the LMs with experts' insights in §3, 2) a reasoning step for both demonstrations and generation to facilitate skill rating, 3) kNN retrieval of few-shot demonstrations from the expert-annotated

[1] For mindful and confident, rate with `yes` or `no`.

	Contrasting Pairs	kNN	Reas- -oning	Curated Rubric	Overall	Describe	Express	Assert	Reinforce	Negotiate	Mindful	Confident
IMBUE (Our method)	✓	✓	✓	✓	**0.6442**	**0.7104**	0.5797	**0.6715**	**0.6873**	**0.7426**	0.5965	**0.5211**
w/o Contrasting Pairs		✓	✓	✓	0.6248	0.6942	**0.5847**	0.6525	0.6257	0.7216	**0.6159**	0.4791
w/o kNN (few-shot)			✓	✓	0.5843	0.6220	0.5275	0.6425	0.5495	0.6651	0.6091	0.4757
w/o kNN (zero-shot)			✓	✓	0.5756	0.5680	0.5797	0.5764	0.5900	0.6723	0.5381	0.5044
w/o Reasoning				✓	0.5020	0.4552	0.5157	0.5427	0.5830	0.6651	0.5922	0.1602
GPT-4					0.3962	0.4690	0.4458	0.4340	0.4620	0.5018	0.3127	0.1480
Llama-2-70b					0.2117	0.1017	0.1632	0.1888	0.1264	0.1311	0.4384	0.3220
Llama-2-13b					0.2366	0.1613	0.4000	0.2882	0.3175	0.2697	0.1629	0.0566
Llama-2-7b					0.2117	0.1017	0.1632	0.1888	0.1264	0.1311	0.2701	0.1806

Table 1: Skill rating baseline and ablation results. We report macro F1 scores of binary classification of `Strong` vs not `Strong` use of each skill. IMBUE, containing all four components: contrasting pair demonstrations, kNN demonstrations, reasoning step, and curated rubric, achieves the highest macro F1 overall, with significant outperformance on Describe, Assert, Reinforce, Negotiate, and Confident skills. IMBUE outperforms GPT-4 by 24.8%.

data in §3, and 4) contrasting pair demonstrations to help LMs learn nuanced concepts.

Curated rubric. To enhance the model's rating calibration, we incorporate information extracted from expert-written feedback into the rating rubric. We use the expert-written improvement suggestions on `weak` responses as well as on `none` responses (where a skill should be applied but was not). We use DBSCAN (Ester et al., 1996) to cluster these improvement suggestions. We then summarize these clusters for each skill and integrate them into the system prompts as an additional rating rubric(Appendix H).

Reasoning step. We follow previous work using chain-of-thought prompting (Wei et al., 2022) to generate the reasoning of a rating before assigning it. We convert expert-written suggestions into the reasoning of ratings and use them as demonstrations. e.g., a suggestion "don't mix feelings and facts" is converted into a reason "the utterance mixes feelings and facts." We perform the conversion using few-shot learning and qualitatively evaluate the conversion with a random sample of 50.

kNN demonstrations. Retrieval-based in-context learning has shown superior performance to comparable approaches in similar tasks (Sharma et al., 2023b). We adapt this approach and retrieve a set of examples from all levels of skill use. We first encode all utterances using the `all-mpnet-base-v2` model with SentenceTransformer. For each query utterance, we use `faiss` (Douze et al., 2024) to retrieve the k most similar examples from each level (`strong`, `weak`, `none`) for this skill in our datasets.

Contrasting pair demonstrations. Utterances often involve the use of multiple skills, posing a challenge for models to identify the text corresponding to each skill. To address this challenge, we construct pairs of (`strong`, `weak`) and (`strong`, `none`) demonstrations. We first search for the k `weak` and `none` examples that are most relevant to the query utterance. We then use the expert rewritten responses as `strong` examples to form the contrasting pairs and use these pairs as demonstrations, which helps the model learn nuanced concepts and disentangle multiple skills. For instance, in the utterance: "*In recent team meetings, my ideas were presented as yours* (Strong Describe)... *this situation has been causing some discomfort* (Strong Express)." Without contrasting pair demonstrations, a model misclassifies it as Weak Describe, suggesting a mixture of facts and feelings. Classifying skill use as weak would trigger unnecessary, if not confusing feedback. However, the underlined subspan corresponding to Describe remains focused on facts, qualifying it as a Strong Describe. We demonstrate empirically that a contrasting pair prompting strategy improves skill rating prediction and the quality of improvement suggestions (§5).

4.2 Next skill suggestion

Before a participant writes utterance U_i, we aim to suggest the set of best skills to use, given situation S and the previous simulated partner's response P_{i-1}. In our dataset, skill L_j is considered "recommended" if: 1) L_j is selected by the participant and the expert does not advise against it, or 2) L_j is not selected but is suggested by the expert. Based on insights from §2, we design the model to always suggest *describe* as the first skill (when $i = 0$). For $i >= 1$, we retrieve the k most similar examples to S concatenated with P_{i-1}, to prompt GPT-4 and generate the suggested skill.

	Human	R-L	BScore	Spec.	Act.
IMBUE	**83%**	10.3	**85.2**	**4.05**	**4.04**
w/o Contrasting Pair	51%	**11.2**	85.0	4.01	3.87
w/o kNN (few-shot)	23%	9.0	84.3	2.55	2.38
w/o kNN (zero-shot)	37%	8.7	84.0	3.75	2.54
w/o Reasoning	34%	7.7	83.8	3.18	2.51
GPT-4	6%	5.8	81.0	1.25	2.28
DEAR MAN Experts	100%	100	100	3.02	4.01

Table 2: Similarity between generated vs experts' improvement suggestions. IMBUE achieves competitive R-L and BScore and the best human evaluation performance. 83% of the time, IMBUE is essentially suggesting the same improvements as DEARMAN experts, based on human eval. Note that automatic metrics should be interpreted with caution, as the gaps in human evaluations are significantly larger. IMBUE achieves highest *specificity* and *actionability*, providing specific and actionable improvement suggestions even more than expert-written suggestions. R-L: ROUGE-L; BScore: BertScore; Spec.: Specificity; Act.: Actionability

5 Evaluation of IMBUE with an Expert-Annotated Dataset

We evaluate IMBUE on the expert-annotated dataset (§3) with cross-validation. We use IMBUE to generate skill use feedback (§4.1) and next skill suggestions (§4.2) and report the average performance across all conversations[2].

Since there are no established methods for the proposed new tasks, we use GPT-4 and Llama-2 variants as a baseline and conduct the following ablations to assess the impact of each component in IMBUE (Tables 1&2): (1) without contrasting pairs component (retrieval-based in-context learning, with reasoning step and curated rubric), (2) without contrasting pairs or kNN-retrieval component (random in-context examples), with reasoning step and curated rubric (3) without in-context examples, only reasoning step and curated rubric, (4) without in-context examples or reasoning, only curated rubrics.

Skill ratings. To maximize feedback opportunities, we prioritize identifying the distinction between `strong` vs. not `strong` skill use, which determines whether the model will provide improvement suggestions. As shown in Table 1, IMBUE achieves the highest macro F1 on average across skills, outperforming GPT-4 by 24.8%. IMBUE outperforms GPT-4 and Llama-2 baselines and all ablations on five out of seven DEAR MAN skills.

[2]To ensure deterministic skill rating predictions, we use `gpt-4-1106-preview` with `temp=0` throughout this paper.

	Macro F1	Entropy
kNN few-shot	**0.5849**	2.23
Random few-shot	0.5723	1.55
Zero-shot	0.5345	1.59
Always suggest the most frequent (Assert)	0.4780	0.00
Random suggestion	0.3899	**2.29**

Table 3: Next skill suggestion, evaluation with expert-annotated dataset. IMBUE gives diverse skill recommendations, almost retaining max entropy (uniform random suggestion) while improving 9% over the second best method's F1 score without kNN demonstrations.

Improvement suggestions. We compare the generated improvement suggestions with expert-written suggestions through a combination of human and automatic evaluation. In our human evaluation, we recruit CS PhD students with significant expertise in NLP and text annotation tasks and ask them to annotate if the expert and model-generated improvement suggestions are similar (on a random sample of 210 pairs; details in §E). We find that IMBUE significantly outperforms baseline and ablations, generating improvement suggestions similar to experts 83% of the time, which is 32% better than the second best. Moreover, we conduct a secondary evaluation using automatic metrics, ROUGE-L (Lin, 2004) and BertScore (Zhang et al., 2019) and find that our model is competitive on both metrics, compared with baseline and ablations. Note that given the open-ended nature of improvement suggestions, automatic metrics are often limited in their ability to capture nuances in what should be considered similar, often focusing on the semantic and linguistic similarity instead of the similarity of the underlying feedback.

We also evaluate the *specificity* and *actionability* of the improvement suggestions. Prior work in NLP to support mental health skills has suggested that feedback that is *specific* to a situation and proposes concrete *actions* is highly preferred and more effective (Sharma et al., 2023b). Here, we use a simple GPT-4-based few-shot prompting method (Ziems et al., 2023a) to measure specificity and actionability. IMBUE outperforms baseline and ablations in both measures, even more than experts, who might be too busy to consistently write highly specific feedback. IMBUE is comparable with experts in *actionablity*, significantly outperforming all baseline and ablations.

Next skill suggestion performance and diversity. To ensure that users receive a diverse range of skill suggestions for practice, we evaluate both the performance of predicting "expert-recommended"

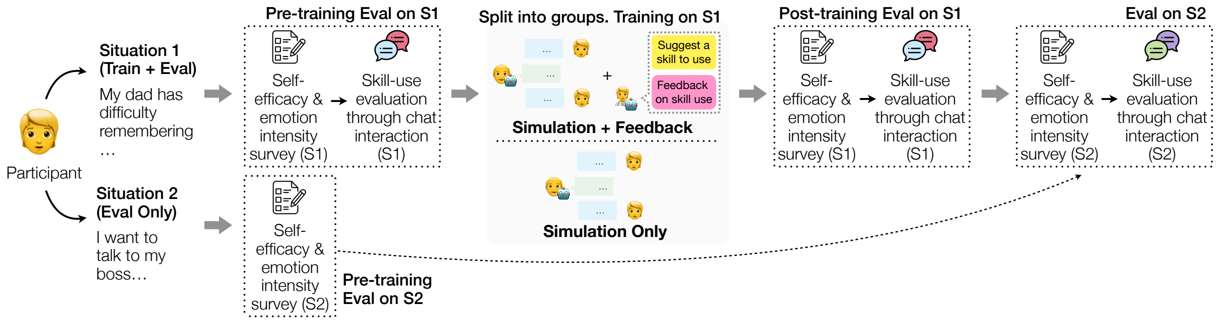


Figure 2: User study experimental design. We randomly assigned participants to one of the simulation-only and simulation+feedback groups. Each participant was asked to provide two situations, S1 and S2. Only S1 was used in training. Both S1 and S2 were used in pre- and post-training self-efficacy and emotion intensity surveys and in post-training skill-use evaluation through chat interaction.

skills and the diversity of the skill suggestions through entropy. As Table 3 shows, our method surpasses the second-best baseline by 9.4% in performance with almost maximum entropy[3].

6 Evaluate IMBUE in a Randomized Trial

We conduct a randomized trial with 86 participants and assess how IMBUE can help people improve interpersonal effectiveness.

6.1 Participant Training Methods

We evaluate two variants for training participants on interpersonal effectiveness – (1) Simulation Only (S) and (2) Simulation + Feedback (S + F).

(1) Simulation Only (S). We develop a GPT-4-based role-playing chatbot designed for participants to converse about their situation (e.g., a chatbot role-playing as the participant's boss). The role-playing chatbot leverages the situation to create a system prompt for GPT-4 (§D). Also, it is designed to be difficult to convince and respond at lengths similar to the length of the participant's message. We qualitatively evaluate this chatbot during our formative study (§2) and data collection (§3). Participants interact with the chatbot to *simulate* the conversation.

(2) Simulation + Feedback (S+F). Using the model developed in §4, we generate the following types of interactive feedback for participants: (1) get a skill suggestion (§4.2), (2) select a skill (can be different from what is suggested) and write a response implementing this skill, (3) get feedback (ratings + improvement suggestions) on skill use (§4.1), (4) improve the response based on the feedback. Steps (2)-(4) can be optionally repeated.

Participants receive this feedback while interacting with the role-playing chatbot designed above to simulate the situation.

To compare IMBUE with current at-home practice, participants in both variants get the DEAR MAN worksheet from the official DBT manual (Linehan, 2014), mirroring current practice.

6.2 Study Procedure and Evaluation Metrics

Figure 2 outlines our study procedure. We recruit participants from mTurk ($n = 34$) and Prolific ($n = 52$). Each participant is asked to provide two difficult communication situations (S1 and S2). Next, they are randomly assigned either Simulation Only (S) or Simulation Feedback (S+F). More details about the study interface procedure are in §J.

The DEAR MAN manual suggests that people's primary struggles in challenging situations are: lack of skills, interference of strong emotions, and fear of not having a successful conversation (Linehan, 2014). Here, we measure the improvement in DEAR MAN **skill mastery**, **emotion reduction** towards the challenging situation, and **self-efficacy** towards having these conversations. We evaluate them *pre-* and *post*-training, enabling a within-person control setup. We also compare the differences between the S and S+F groups, distilling the effect of just-in-time feedback.

6.3 RQ1: How do simulation and feedback improve DEAR MAN skill mastery?

We measure a participant's skill mastery with our model in §4 before and after the training. We then compare the level of skill use in *pre-* and *post*-training evaluation chat to evaluate the effect of training on the situation that the participant is being trained on (S1). Also, to test the generalizability of the skill learning, we also conduct evaluation on a new and more difficult situation (S2) in which the

[3] IMBUE can recommend multiple skills at the same time, here we evaluate on single skill recommendation, so users can focus on improving one skill at a time.

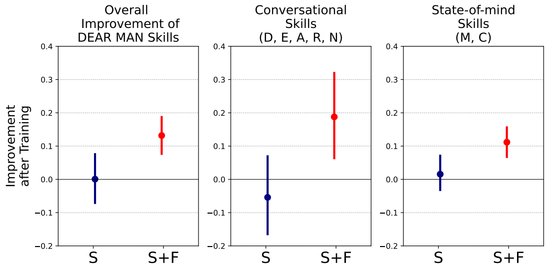


Figure 3: Improvement in skill mastery. Simulation+feedback group shows a significantly higher improvement in skill mastery (17.6% on a 0-2 scale, **, d=0.59) compared to simulation-only (0.1%) after only one training session. The difference is also significant for the subset of conversational skills that participants choose to use in each utterance (only measured when the skills are chosen), *Describe*, *Express*, *Assert*, *Reinforce*, and *Negotiate* (24.8%, **, d=0.59) and state-of-mind skills (measured in every utterance), *Mindful* and *Confident* (15.7%, **, d=0.59). (***: $p < .001$, **:$p < .01$, *:$p < .05$. d: Cohen's d.)

participant has not been trained on and does not receive feedback during conversation simulation.

Figure 3 compares S and S+F groups on the improvement of skill mastery for S1, the situation used during training. S+F group shows a significantly higher improvement in skill mastery after one training conversation by 17.6% ($p = .007$, Cohen's $d = 0.59$) on a score scale of 0-2, compared to the S group which improved by only 0.1%. This difference is also significant for the set of five conversation skills, *Describe*, *Express*, *Assert*, *Reinforce*, and *Negotiate*, at 24.8% ($p = .008, d = 0.59$) and the set of state-of-mind skills, *Mindful* and *Confident*, at 15.7% ($p = .007, d = 0.59$). Among all the skills, S+F shows significant more improvement in *Express*, *Mindful*, and *Confident* (Appendix Figure 5).

6.4 RQ2: How do simulation and feedback enhance emotion reduction?

We evaluate emotion reduction on four negative emotions from the Plutchik's Wheel (Plutchik, 1980): anger, fear, sadness, disgust. We ask the participants rate their agreement to statements like "I feel sad about the situation" using a 7-point Likert scale (Likert, 1932) before and after the training. S+F group shows significant reduction of almost all negative emotions on S1. We find that both S group and S+F group have reduced **fear** (by 25.1%, 40.8%, $p = .000, .000$, $d = .71, 1.19$) and **sadness** (by 17.3%, 29.9%, $p = .020, .000$, $d = .45, .76$) towards the situation after training. S+F group shows a significantly higher reduction towards **fear** (by 15.7%, $p = .021$, $d = .51$), compared to S group. S+F group also has a significant reduction in **anger** (by 23.5%, $p = 0.030$), whereas the S group does not show significant change.

6.5 RQ3: How do simulation and feedback improve participants' self-efficacy?

To evaluate participants' self-efficacy before and after the training, we ask the participants to rate their confidence, worry, hopefulness, and motivation about having the conversation before and after the training, again with a 7-point Likert-scale.

As Figure 4 shows, both S and S+F groups show a significant increase in self-reported **confidence** (by 16.9%, 43.6%, $p = .035, .000$, $d = .41, 1.08$) and a significant reduction in self-reported **worry** (by 26.9%, 30.9%, $p = .000, .000$, $d = .81, 1.04$) towards having the conversation. Moreover, S+F group demonstrates significantly higher increases in **confidence** (by 26.7%, $p = .010$, $d = .57$), compared to the S group, underscoring the effectiveness of just-in-time feedback. In addition, S+F group showed a significant increase in **hopefulness** (by 11.0%, $p = .046$, $d = .35$) and **motivation** (by 22.1%, $p = .001$, $d = .62$) towards having the conversation, whereas the S group does not show significant change. This shows that S+F version of the tool helps in these dimensions, though we cannot separate the effect of just-in-time feedback and repeated practice with the simulation.

6.6 RQ4: Do these effects generalize to a new and more difficult situation?

We compare the average skill used in conversation on S2 and both *pre*- and *post*-training evaluation conversation on S1. We find that for S+F group, the average skill use rating is significantly higher in conversation on S2 compared to *pre*-training conversation on S1 ($p = .049$). The skill use ratings are not significantly different between conversation on S2 and *post*-training conversation on S1. These comparisons show that the skill use improvement can be generalized, without significant diminishing effect, to a new and more difficult situation immediately after training.

Although skill mastery generalizes to a new and more difficult situation (S2), self-efficacy and emotional reduction do not immediately generalize (Figure 4). Many constructs, such as confidence, hopeful, worry, fear and sadness, show a positive improvement but these differences were not statistically significant at $\alpha = 0.05$. This could be

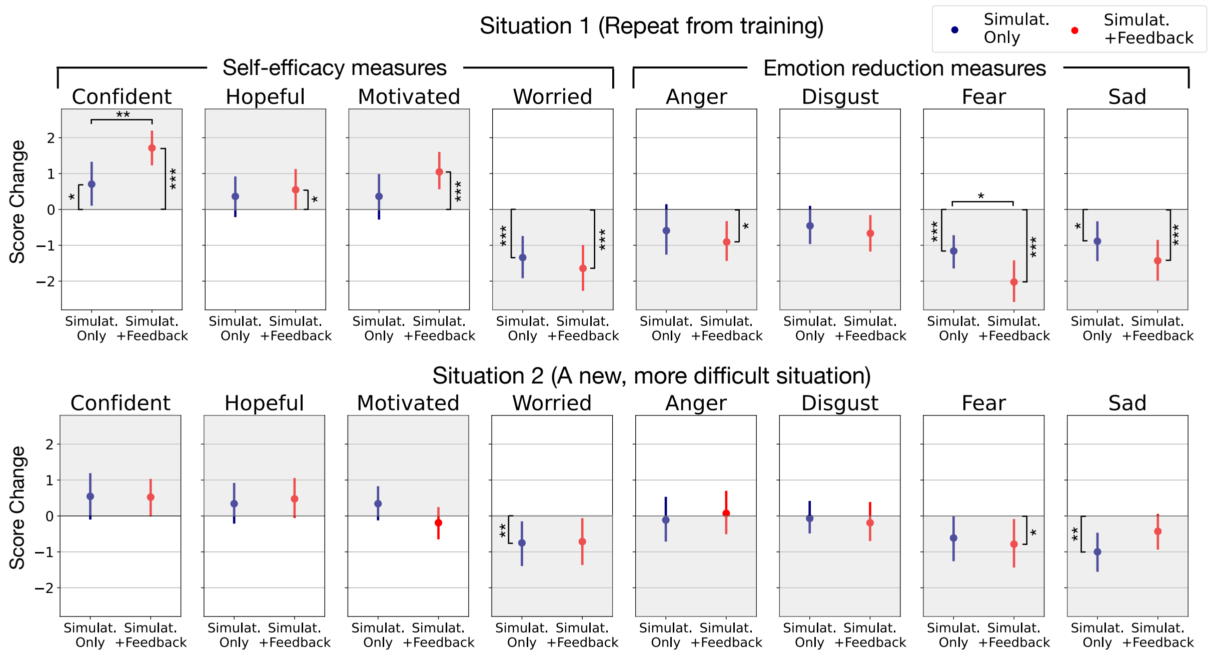


Figure 4: Change of self-reported efficacy and emotional intensity for both Situation 1 (S1) and Situation 2 (S2) after a single training session on S1. Gray area indicates the direction of improvement for each score. The group receiving just-in-time feedback generated with our method in addition to conversation simulation see significant increase in their confidence (43.6%, ***), hopefulness (11.0%, *), motivation (22.1%, ***) towards having the conversation, significant decrease in their worrying thoughts (30.9%, ***) about having the conversation and their anger (23.5%, *), fear (40.9%, ***), and sadness (29.0%, ***) towards the training situation (S1). The increase in confidence and reduction in fear are 26.7% (**, d=0.57) and 15.7% (*,d=0.51) significantly more than the group receiving simulation only. This improvement in self-efficacy and emotional reduction does not transfer immediately to a new, more difficult situation (S2). See Section 6 for more analysis and discussion.

attributed to the difficulty of managing emotions in novel situations without specific training, suggesting that targeted emotional regulation training of a different type or over an extended period may be necessary (Freitas and Salovey, 2000).

The findings also emphasize that practicing in simulations with feedback tailored to the exact situation is more effective for improving self-efficacy and managing emotions. Our tool supports exactly this accessibility, lowering the barrier to effective learning and practices.

7 Related Work

NLP literature. Broadly, our work is related to the growing body of works on LLM-based autonomous agents (Park et al., 2022, 2023; Argyle et al., 2023b,a; Zhou et al., 2023b; Liu et al., 2023a; Aher et al., 2022; Wang et al., 2023; Dubois et al., 2023) and using LLM in psychology and computational social science (Ziems et al., 2023b; Demszky et al., 2023; Sharma et al., 2023b,a; Lin et al., 2022; Pérez-Rosas et al., 2022; Shah et al., 2022; Sharma et al., 2020a,b; Wadden et al., 2021; Welch et al., 2020; Zhang and Danescu-Niculescu-Mizil, 2020; Gaur et al., 2019; Lee et al., 2019; Pérez-Rosas et al., 2019; Althoff et al., 2016). Our work most closely relates to recent works using LMs in their roleplaying capacity to facilitate communication skill learning (Shaikh et al., 2023; Liu et al., 2023b; Argyle et al., 2023a). Our work is the first to focus on both communication skills and emotion management simultaneously, incorporate experts' domain knowledge in providing feedback, and ground in clinical psychology theory.

Psychology literature. Dialectical Behavior Therapy (DBT) is a well-established treatment program that has been identified as having strong efficacy based on several Randomized Control Trials (RCTs) (Panos et al., 2014; Oldham, 2005). DBT contains several components including DEAR MAN, emotional regulation, and other skills. It is common practice for a client to receive training for DBT as a whole, rather than specific components, although therapists may emphasize certain components over others based on the client's specific situation. To the best of our knowledge, there are no reports on the specific effect of training on DEAR MAN. The APA Practice Guidelines also point out that "It is difficult to ascertain whether the improvement reported for patients receiving DBT derived from specific ingredients of DBT." (Association et al., 2002) The lack of granular evaluation on training programs is a known and common problem in the psychology field beyond DBT (Alexopoulos and Arean, 2014).

Our work complements the current practice of DBT, specifically using LMs' simulation capability to offer more accessible learning and practice opportunities for DEAR MAN skills outside of treatment sessions. In our randomized trial, both the Simulation and Simulation+Feedback groups demonstrate significant improvements compared to the "pure control," which represents participants' performance in pre-training conversations. The Simulation mirrors in-person practices where therapists take on the role of the conversation partner. Therefore, the effectiveness demonstrated in the Simulation-only group partially addresses the absence of DEAR MAN-specific efficacy evaluation, contributing to the existing literature.

8 Conclusion

In this paper, we demonstrate how human-LM interaction can be used to facilitate interpersonal effectiveness skill learning and practice. We collect a dataset with crowd workers and clinical experts who are specifically trained in the practice of DBT. Using this dataset, we develop methods to prompt LMs to simulate bespoke communication scenarios and provide just-in-time feedback, grounded in psychotherapy theory. We build an interactive training system IMBUE, and conduct a randomized user study with 86 participants to assess the effectiveness of the simulation and feedback components of IMBUE. We find that simulation-only training is effective in improving self-efficacy and emotion reduction, and adding just-in-time feedback shows significantly more benefits in all of skill mastery, self-efficacy, and emotion reduction. The skill mastery can be acquired from practicing with a different situation, while emotion reduction and self-efficacy appear to only benefit from training specifically on the situation.

Ethics Statement

IRB Approval. We obtained approval from our institution's Institutional Review Board (IRB) for our study. Our institution requires all researchers who conduct human subjects research to complete human subjects protection training. The researchers who conducted this study were certified by the IRB.

Informed Consent from Participants. We obtained consent from participants in both our data collection and the user study. All participants were aged 18 and older. Participants were informed that they were interacting with an AI-based model simulating their conversation partner and the data they provided would be released for research purposes. Participants were also informed that some content from the model might be upsetting since the conversation might get heated.

Crisis Resources. We use an API with content filters to minimize the possibility of harmful output during deployment. [4] Nevertheless, some content might still be upsetting to the participants. We provide two crisis resources to our participants during the study: Crisis Text Line (crisistextline.org) and 988 Suicide and Crisis Lifeline (988lifeline.org). We did not observe any adverse events.

Privacy. Our study does not collect Privately Identifiable Information (PII) and we asked that participants avoid including any PII in the situations or conversations. The conversations and situations were manually filtered to ensure there were no identifiable names or locations.

Limitations

Our work is not without limitations. Importantly, we note that our tool is not meant to replace practice with an expert. Rather, we built the tool to complement current practice and to lower the barrier of access to learning and practicing. We note further limitations below.

We do not directly address the potential for dual-use of our tool, especially if it is used by individuals with harmful intentions towards others. However, we prioritize improving the well-being and mindfulness of conversation participants rather than simply helping them win negotiations. For example, we consider it a suboptimal case if someone "wins" a negotiation but is not being mindful and has negative emotional swings during the process. This consideration is based on insights from experts in §2. By focusing on mindfulness and emotional well-being, we aim to mitigate the dual-use dilemma.

We do not assist participants in setting goals. In our randomized trial, we choose participants who

[4] https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter

can clearly express their goals and work with them to achieve these goals. Setting the right goal is crucial but can be challenging, and other frameworks in DBT address this issue. We have begun by collecting goals and expert annotations in our data collection for future research to expand upon.

Due to the design of our study, the study length is already about an hour. To avoid cognitively overloading the participants, we asked them to do only one training session. We did not investigate the effect of different "dosages" of training. In addition, short-term improvement may not imply long-term improvement, further work is needed to investigate the long-term effect of using such a tool. However, we note that a key benefit of our system is the just-in-time availability, which allows practice just before the user anticipates a challenging conversation.

To minimize participant burden, we collect self-reported scores for emotion reduction and self-efficacy constructs through single questions, rather than a comprehensive survey. We use common measures like "sad" and "angry" for emotions and like "confident" and "worry" for self-efficacy to prevent reporting biases due to misinterpretation. However, it is important to note that self-reported scores, while commonly used in mental health assessments, may contain biases and inaccuracies (Stone et al., 1999).

Our experimental design does not consider individuals with specific mental health conditions that could impact communication. Additionally, we do not address cultural variations in communication, recognizing that what may be perceived as confidence in one culture could be seen as aggression in another. We leave it to future work to develop more personalized and culturally sensitive communication training tools.

Language models and data annotations are known to contain biases (Santurkar et al., 2023; Zhou et al., 2023a; Durmus et al., 2023; Lin et al., 2022; Aguirre et al., 2023; Hovy and Prabhumoye, 2021). In our context, the simulation step may contain persona bias (Gupta et al., 2024). Our tool, designed with *Insight 1* in §2, steers participants to focus on facts and avoids characterizing personalities. This mitigates the risk of triggering the LM to exhibit persona biases. Nonetheless, a thorough assessment of bias and safety is necessary before deploying a tool of this nature in the real world.

Acknowledgments

We are grateful to the clinical experts for their insights during the formative study and data annotation, as well as to the participants from Prolific and mTurk for their participation and feedback. We appreciate the feedback on interface development and design from Kaiming Cheng, Mike Merrill, and Galen Weld. Additionally, we extend our thanks to Bret Nestor, Ken Gu, Margaret Li, Oscar Liu, Vinayak Gupta, and Xinyi Zhou for their suggestions and feedback throughout the course of this project.

I.W.L., A.S., C.M.R., and T.A. were supported in part by the Washington Research Foundation, NSF CAREER grant IIS-2142794, NSF grant IIS-1901386, NIH grant R01MH125179, Bill & Melinda Gates Foundation (INV-004841), and the Office of Naval Research (#N00014-21-1-2154).

References

Carlos Aguirre, Kuleen Sasse, Isabel Cachola, and Mark Dredze. 2023. Selecting shots for demographic fairness in few-shot learning with large language models. *arXiv preprint arXiv:2311.08472*.

Gati Aher, RosaI. Arriaga, and Adam Tauman Kalai. 2022. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*.

George S Alexopoulos and Patricia Arean. 2014. A model for streamlining psychotherapy in the rdoc era: the example of 'engage'. *Molecular psychiatry*, 19(1):14–19.

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

Lisa P Argyle, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023a. Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023b. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

American Psychiatric Association et al. 2002. *Practice guideline for the treatment of patients with bipolar disorder (revision)*. American Psychiatric Pub.

Aaron T Beck. 1979. *Cognitive therapy and the emotional disorders*. Penguin.

Aaron T Beck. 1996. Beyond belief: A theory of modes, personality, and psychopathology.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron A. Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2:688 – 701.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. *ArXiv*, abs/2305.14387.

Esin Durmus, Karina Nyugen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *ArXiv*, abs/2306.16388.

Martin Ester, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

Antonio Freitas and Peter Salovey. 2000. Regulating emotion in the short and long term. *Psychological Inquiry*, 11(3):178–179.

Robert M Gagne. 1965. The analysis of instructional objectives for the design of instruction. *Teaching machines and programmed learning II: Data and directions*, pages 21–65.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *WWW*.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias Runs Deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Peter Hartley. 2002. *Interpersonal communication*. Routledge.

Fiona A L Henderson. 2016. *Difficult conversations on the frontline. Managing the tensions between care and control: are communication skills enough?* Ph.D. thesis, University of Essex.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Hanna Kallio, Anna-Maija Pietilä, Martin Johnson, and Mari Kangasniemi. 2016. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of advanced nursing*, 72(12):2954–2965.

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*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. Gendered mental health stigma in masked language models. In *EMNLP*.

Marsha Linehan. 2014. *DBT Skills training manual*. Guilford Publications.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. 2023a. Training socially aligned language models in simulated human society. *ArXiv*, abs/2305.16960.

Ryan Liu, Howard Yen, Raja Marjieh, Thomas L Griffiths, and Ranjay Krishna. 2023b. Improving interpersonal communication by simulating audiences with language models. *arXiv preprint arXiv:2311.00687*.

Donna Luff, Elliott B. Martin, Kelsey Mills, Natalia M. Mazzola, Sigall K. Bell, and Elaine C. Meyer. 2016. Clinicians' strategies for managing their emotions during difficult healthcare conversations. *Patient Education and Counseling*, 99(9):1461–1466. Communication in Healthcare: Best papers from the International Conference on Communication in Healthcare, New Orleans, LA, USA, October 25-28, 2015.

John M Oldham. 2005. Guideline watch: practice guideline for the treatment of patients with borderline personality disorder. *Focus*, 3(3):396–400.

Mark Olfson. 2016. Building the mental health workforce capacity needed to treat adults with serious mental illnesses. *Health Affairs*, 35(6):983–990. PMID: 27269013.

Patrick T Panos, John W Jackson, Omar Hasan, and Angelea Panos. 2014. Meta-analysis and systematic review assessing the efficacy of dialectical behavior therapy (dbt). *Research on social work practice*, 24(2):213–223.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Verónica Pérez-Rosas, Kenneth Resnicow, Rada Mihalcea, et al. 2022. Pair: Prompt-aware margin ranking for counselor reflection scoring in motivational interviewing. In *EMNLP*.

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*.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion.

Marshall B Rosenberg and Deepak Chopra. 2015. *Nonviolent communication: A language of life: Life-changing tools for healthy relationships*. PuddleDancer Press.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? *ArXiv*, abs/2303.17548.

Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2022. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *CSCW*.

Omar Shaikh, Valentino Chai, Michele J Gelfand, Diyi Yang, and Michael S Bernstein. 2023. Rehearsal: Simulating conflict to teach conflict resolution. *arXiv preprint arXiv:2309.12309*.

Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020a. Engagement patterns of peer-to-peer interactions on mental health platforms. In *ICWSM*.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023a. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*.

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

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G Lucas, Adam S Miner, Theresa Nguyen, and Tim Althoff. 2023b. Cognitive reframing of negative thoughts through human-language model interaction. *arXiv preprint arXiv:2305.02466*.

Arthur A Stone, Christine A Bachrach, Jared B Jobe, Howard S Kurtzman, and Virginia S Cain. 1999. *The science of self-report: Implications for research and practice*. Psychology Press.

Douglas Stone, Bruce Patton, and Sheila Heen. 2023. *Difficult conversations: How to discuss what matters most*. Penguin.

David Wadden, Tal August, Qisheng Li, and Tim Althoff. 2021. The effect of moderation on online mental health conversations. In *ICWSM*.

Lei Wang, Chengbang Ma, Xueyang Feng, Zeyu Zhang, Hao ran Yang, Jingsen Zhang, Zhi-Yang Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji rong Wen. 2023. A survey on large language model based autonomous agents. *ArXiv*, abs/2308.11432.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Charles Welch, Allison Lahnala, Verónica Pérez-Rosas, Siqi Shen, Sarah Seraj, Larry An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2020. Expressive interviewing: A conversational system for coping with covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Andrew Lukas Yin, Pargol Gheissari, Inna Wanyin Lin, Michael Sobolev, John P. Pollak, Curtis Cole, and

Deborah Estrin. 2020. Role of technology in self-assessment and feedback among hospitalist physicians: Semistructured interviews and thematic analysis. *Journal of Medical Internet Research*, 22.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *ACL*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023a. Navigating the grey area: Expressions of overconfidence and uncertainty in language models. *ArXiv*, abs/2302.13439.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2023b. Sotopia: Interactive evaluation for social intelligence in language agents. *ArXiv*, abs/2310.11667.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023a. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

Caleb Ziems, William B. Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023b. Can large language models transform computational social science? *ArXiv*, abs/2305.03514.

A DEAR MAN Definition (Linehan, 2014)

DEAR MAN Skill	Definition and Examples
Describe	Describe the current situation (if necessary). Stick to the facts. Tell the person exactly what you are reacting to. e.g. You told me you would be home by dinner but you didn't get here until 11.
Express	Express your feelings and opinions about the situation. Don't assume that the other person knows how you feel. e.g. When you come home so late, I start worrying about you.
Assert	Assert yourself by asking for what you want or saying no clearly. Do not assume that others will figure out what you want. Remember that others cannot read your mind. e.g. I would really like it if you would call me when you are going to be late.
Reinforce	Reinforce the person ahead of time by explaining the positive effects of getting what you want or need. If necessary, also clarify the negative consequences of not getting what you want or need. e.g. I would be so relieved, and a lot easier to live with, if you do that.
Mindful	Keep your focus on your goals. Maintain your position. Don't be distracted. Don't get off the topic. e.g. I would still like a call
Appear Confident	Appear effective and competent. Use a confident voice tone. Avoid saying things like "I'm not sure."
Negotiate	Be willing to give to get. Offer and ask for other solutions to the problem. Reduce your request. Say no, but offer to do something else or to solve the problem another way. Focus on what will work. e.g. How about if you text me when you think you might be late?

B Formative Study Details

We recruit from the clinical psychology departments in four universities and select those who indicated in the signup form that they "sometimes" or "regularly" work with clients on DEAR MAN skills(§C). We conduct the study with three clinical experts (E1, E2, E3) in semi-structured interviews (Kallio et al., 2016; Yin et al., 2020). We first show the experts a preliminary version of the interface, of which the main functions include: collecting information from users about a difficult situation they face and an LM-backed chatbot that is instructed to roleplay the conversation partner in the user-specified situation who is difficult to convince. We first let the expert try the interface, followed by structured questions on skill teaching, learning, and practice, and the measurement of success in DEAR MAN skill acquisition. We share the insights that informed us about several design decisions. Clinical experts are paid $37.5/hour for this two-hour task.

C Expert Recruitment - DEAR MAN experience question

In the expert signup form, we specifically ask for their experience with DBT and DEAR MAN skills. We only selected those who chose "5 - I sometimes work with clients on DEAR MAN in my practice" or "6 - I regularly work with clients on DEAR MAN in my practice".

- 1 - I have only heard about it
- 2 - I have learned about it in school / read about it extensively but never used it in practice
- 3 - I have worked with clients on DBT but not DEAR MAN specifically
- 4 - I have worked with clients on DEAR MAN at least once
- 5 - I sometimes work with clients on DEAR MAN in my practice
- 6 - I regularly work with clients on DERA MAN in my practice

D Simulation - System Prompt

We use the below prompt as an input to an LM, to generate a system prompt for the simulation LM.

"Situation: My husband always comes home late and he doesn't text me or call me. Prompt: Act like my husband who always comes home late without calling or texting me. Prompt: Act like my boss who regularly calls me on weekends but I don't want to work on the weekends. Situation: My friend has depression and she relies on me 24/7 and I feel drained. Prompt: Act like my friend who has depression and who relies on me whenever you have an issue and I want to convince you to seek professional help and not rely on a friend for all your issues. Situation: My neighbor frequently plays loud music at a late hour and hosts big parties, which affect my sleep. Prompt: Act like my neighbor. You frequently play loud music at a late hour and host big parties. Situation: The airline lost my luggage and the customer service agents have been passing the buck. Prompt: Act like a customer service agent. Your airline lost my luggage and your colleagues have been passing the buck."

(We use \n in the prompt separating each situation.)

E Human Evaluation

We conducted a human evaluation with 13 CS PhD students who are familiar with LM generations. The task is to identify if the model-generated feedback is the same as expert-generated feedback. The participants are guided by the rubric below:

1). Think about if you received the feedback, would you have acted the same, or differently?
2). Do they give the same *reason* of why the utterance should be improved?

3). Do not think about whether they are equally specific or actionable. (We account for these two measures separately.)
4). If one of the feedback seems to be noisy data, like if they just output a single word "Describe", they are not considered the same.

The participants discussed a few examples, like the following, at the beginning to help guide the annotation:

Example 1
Expert: Try using more non-judgmental language here and focus on describing the details - think about what a video camera would see.
Model: To improve the utterance should focus on describing the actions without assigning motive or making judgmental statements For example "You expressed dissatisfaction with the gift from your in-laws in public" would be more objective
Label: Same.
Example 2
Expert: Staying mindful means holding your ground and avoiding being distracted by the server's comments about how good the other beer tastes.
Model: To align with the DEAR MAN framework the utterance could be structured to Describe the situation Express feelings Assert needs Reinforce the positive stay Mindful Appear confident and Negotiate if necessary For example
Label: Not the same

The participants are unaware of which source model their data originates from, and none of them have seen any generations from any of the model variants prior to evaluation. Each pair of model-generated and expert feedback receives three annotations. We consider the majority of these three annotations as the final annotations. The average Fleiss' Kappa of the annotations is: 0.667, showing substantial agreement between the annotators.

F Specificity and Actionability - Few shot learning prompts

Actionability

On a scale from 1 to 5, how actionable is the following feedback?

Feedback: Stay on topic and approach the situation with the intention of finding a resolution Consider expressing your concerns and the impact of your mother's actions more calmly and objectively rather than accusing her of negative intentions Actionability: 4

Feedback: To align with the DEAR MAN framework the utterance could be more assertive and clear about the need and the reason behind the request For example Actionability: 3

Feedback: I would suggest using express more heavily to express appreciation and understanding of the other person's point of view. Actionability: 3

Feedback: To align the utterance with the DEAR MAN framework which is a skill from Dialectical Behavior Therapy (DBT) used to teach effective communication the speaker could structure their statement with more clarity and respect focusing on the following components Actionability: 2

Feedback: You can try phrasing your ask more assertively, using "I want" rather than "should". Actionability: 5

Feedback: 1 **Describe** Actionability: 1

Feedback: To align the utterance more closely with the DEAR MAN framework which is a skill from Dialectical Behavior Therapy (DBT) designed to help people communicate effectively and assertively the utterance could be structured as follows Actionability: 1

Specificity

On a scale from 1 to 5, how specific is the following feedback, given the situation and the utterance?

Situation: My colleague keeps borrowing money from me without completely paying her old debts, and she doesn't feel ashamed to come asking despite I've confronted her several times about it. But I don't know her to refuse to lend to her because I have it and can't lie I don't have any money Utterance: I'm not happy with you any longer because you owe me a lot of money promising to pay it back soon but you haven't. Feedback: Be more specific with how much money, and how often, your friend is borrowing. Specificity: 5

Situation: my friend borrowed my car and hid the fact that he had a slight accident and dented my car Utterance: oh yes you did and you know it, I already found out the truth because Jenny told me what you guys did the other day, but here you are lying to my face Feedback: This utterance demonstrates mindfulness by focusing on the issue of the car accident and the dishonesty rather than getting sidetracked by other topics It's direct and addresses the core issue effectively Specificity: 4

Situation: I went to dinner with my friends and a waiter brought me the wrong beer for the second time. I had asked for a Blue Moon but they kept bringing me Samuel Adams. Utterance: No worries. Why no Blue Moon? I'm just curious. Feedback: The speaker maintains composure and expresses curiosity rather than frustration or anger indicating mindfulness in addressing the mistake without getting sidetracked by emotions Specificity: 3

Situation: At the library, a guest has the phone on loud and we can hear every time they receive a text. Utterance: But we'll still hear the sound of your incoming texts. Feedback: This utterance is appropriate as it is It objectively describes the situation without adding any unnecessary judgment or emotion Specificity: 2

Situation: My colleague keeps borrowing money from me without completely paying her old debts, and she doesn't feel ashamed to come asking despite I've confronted her several times about it. But I don't know her to refuse to lend to her because I have it and can't lie I don't have any money Utterance: I'm not happy with you any longer because you owe me a lot of money promising to pay it back soon but you haven't. Feedback: To align the utterance more closely with the DEAR MAN framework which is a skill from Dialectical Behavior Therapy (DBT) designed to help people communicate effectively and assertively the utterance could be structured as follows Specificity: 1

G User Study Results

	treatment_post_scores	treatment_pre_scores	perc_t	treatment_t_stat	treatment_p_value	symbol	Cohen's d
Confident	5.64	3.93	44.0%	6.48	0.00	***	1.08
Worried	3.67	5.31	-31.0%	-5.2	0.00	***	-1.04
Hopeful	5.52	4.98	11.0%	2.05	0.05	*	0.35
Motivated	5.79	4.74	22.0%	3.68	0.00	***	0.62
Anger	2.95	3.86	-23.0%	-2.25	0.03	*	-0.47
Fear	2.93	4.95	-41.0%	-6.54	0.00	***	-1.19
Disgust	2.38	3.05	-22.0%	-1.85	0.07		-0.33
Sad	3.5	4.93	-29.0%	-4.77	0.00	***	-0.76

Table 4: User Study Results - Simulation+Feedback. **Situation 1**. Improvement after the training for Treatment and Control groups. Significance means there is a significant increase in self-reported efficacy or emotions after training.

	control_post_scores	control_pre_scores	perc_c	control_t_stat	control_p_value	symbol_c	Cohen's d
Confident	4.86	416.0%	0.17	2.18	0.03	*	0.41
Worried	3.64	498.0%	-0.27	-4.5	0.00	***	-0.81
Hopeful	4.91	455.0%	0.08	1.21	0.23		0.2
Motivated	5.09	473.0%	0.08	1.12	0.27		0.2
Anger	3.61	420.0%	-0.14	-1.53	0.13		-0.32
Fear	3.45	461.0%	-0.25	-3.87	0.00	***	-0.71
Disgust	2.7	316.0%	-0.14	-1.42	0.16		-0.27
Sad	4.25	514.0%	-0.17	-2.42	0.02	*	-0.45

Table 5: User Study Results - Simulation-Only. **Situation 1**. Improvement after the training for Treatment and Control groups. Significance means there is a significant increase in self-reported efficacy or emotions after training.

emotion	treatment_post_scores	treatment_pre_scores	perc_t	treatment_t_stat	treatment_p_value	symbol
Confident	4.214	3.690	0.142	1.834	0.074	
Worried	4.714	5.429	-0.132	-1.888	0.066	
Hopeful	4.905	4.429	0.108	1.800	0.079	
Motivated	4.881	5.071	-0.038	-0.840	0.406	
Anger	4.548	4.476	0.016	0.215	0.831	
Fear	4.214	5.000	-0.157	-2.118	0.040	*
Disgust	3.357	3.548	-0.054	-0.797	0.430	
Sad	4.548	4.976	-0.086	-1.232	0.225	

Table 6: User Study Results - Simulation+Feedback. **Situation 2.** Improvement after the training for Treatment and Control groups. Significance means there is a significant increase in self-reported efficacy or emotions after training.

	control_post_scores	control_pre_scores	perc_c	control_t_stat	control_p_value	symbol_c
Confident	4.068	3.523	0.155	1.312	0.196	
Worried	4.545	5.295	-0.142	-2.096	0.042	*
Hopeful	4.341	4.000	0.085	1.106	0.275	
Motivated	4.682	4.341	0.079	0.965	0.340	
Anger	3.955	4.068	-0.028	-0.292	0.772	
Fear	4.273	4.886	-0.126	-1.414	0.165	
Disgust	3.250	3.318	-0.021	-0.230	0.819	
Sad	4.205	5.205	-0.192	-2.819	0.007	**

Table 7: User Study Results - Simulation-only. **Situation 2.** Improvement after the training for Treatment and Control groups. Significance means there is a significant increase in self-reported efficacy or emotions after training.

emotion	treatment_diff	control_diff	effect_size	T-C_t_stat	T-C_p_value	symbol
Confident	1.71	0.70	0.57	2.652	0.010	**
Worried	-1.64	-1.34	-0.15	-0.707	0.482	
Hopeful	0.55	0.36	0.10	0.463	0.644	
Motivated	1.05	0.36	0.36	1.683	0.096	
Anger	-0.90	-0.59	-0.15	-0.710	0.479	
Fear	-2.02	-1.16	-0.51	-2.357	0.021	*
Disgust	-0.67	-0.45	-0.13	-0.586	0.559	
Sad	-1.43	-0.89	-0.30	-1.400	0.165	

Table 8: User Study Results. Difference in difference for Situation 1. The significant result means treatment group and control group are significantly different.

emotion	treatment_diff	control_diff	effect_size	T-C_t_stat	T-C_p_value	symbol
Confident	0.71	1.18	-0.24	-0.686	0.498	
Worried	-1.65	-0.65	-0.50	-1.471	0.151	
Hopeful	0.53	0.41	0.05	0.159	0.874	
Motivated	-0.18	0.47	-0.37	-1.074	0.291	
Anger	-0.12	-0.88	0.39	1.135	0.265	
Fear	-1.53	-0.82	-0.31	-0.902	0.374	
Disgust	-0.35	-0.41	0.03	0.101	0.920	
Sad	-0.59	-1.29	0.38	1.112	0.275	

Table 9: User Study Results. Difference in difference for Situation 2. The significant result means treatment group and control group are significantly different.

H System prompts used in IMBUE

Skill	System Prompt
Describe	You will be given a context and an utterance, from a conversation that happened in the given context. Does the given utterance describe the given context? To be considered "describe", the utterance needs to stick to the facts, make no judgmental statements, and be objective. Rating Rubric: A "Strong Describe" rating indicates that the utterance is or contains a description of the given context. It sticks to the facts, makes no judgemental statements, and is objective. Do ALL of the following three steps. Step 1: Generate "Reasoning for rating". Step 2: Generate "Describe Rating" in "Strong Describe", "Weak Describe" or "No Describe". A "Weak Describe" rating indicates that the utterance is or contains a description of the given context, but needs improvement since it may not stick to the fact, makes some judgemental statements, or is not fully objective. A "No Describe" rating indicates that the utterance does not describe any aspect of the given context at all. Step 3: Provide additional comments on the ratings similar to the examples given. Finish each step with ###. Twenty words minimum. YOU MUST FINISH EACH STEP WITH ###
Express	You will be given a context and an utterance, from a conversation that happened in the given context. Does the given utterance explicitly express how the speaker feels in the conversation? To be considered "express", the utterance needs to EXPLICITLY express your feelings about the given context, including things like "this makes me feel", "I feel ... by your actions", "this situation/your action... has caused me..." with adjectives or nouns describing emotions. Do ALL of the following three steps. Step 1: Generate "Reasoning for rating". Step 2: Generate "Express Rating" in "Strong Express", "Weak Express" or "No Express". Rating Rubric: A "Strong Express" rating indicates that the utterance is or contains an EXPLICIT expression of the felt emotions. YOU CANNOT INTERPRET THE SENTIMENT IF THE SPEAKER DOES NOT MENTION EMOTIONS. A "Weak Express" rating indicates that the utterance is or contains an expression of your feelings or opinions about the given context, but can be made more explicit in expressing feelings. A "No Express" rating indicates that the utterance does not express your feelings or opinions about the given context at all. Step 3: Provide additional comments on the ratings similar to the examples given. Finish each step with ###. Twenty words minimum. YOU MUST FINISH EACH STEP WITH ###
Assert	You will be given a context and an utterance, from a conversation that happened in the given context. Does the given utterance assert your needs or wants about the given context? To be considered "assertive", the utterance needs to be asking for what you want or saying no clearly. Do ALL of the following three steps. Step 1: Generate "Reasoning for rating". Step 2: Generate "Assert Rating" in "Strong Assert", "Weak Assert" or "No Assert". Rating Rubric: A "Strong Assert" rating indicates that the utterance is or contains an assertion of your needs or wants about the given context. A "Weak Assert" rating indicates that the utterance is or contains an assertion of your needs or wants about the given context, but needs improvement in making it more explicit or stronger. A "No Assert" rating indicates that the utterance does not contain an assertion of the needs or wants. Step 3: Provide additional comments on the ratings similar to the examples given. Finish each step with ###. Twenty words minimum. YOU MUST FINISH EACH STEP WITH ###
Reinforce	You will be given a context and an utterance, from a conversation that happened in the given context. Does the given utterance reinforce your needs or wants about the given context? To be considered "reinforce", the utterance needs to reinforce some reward for the other person. Do ALL of the following three steps. Step 1: Generate "Reasoning for rating". Step 2: Generate "Reinforce Rating" in "Strong Reinforce", "Weak Reinforce" or "No Reinforce". Rating Rubric: A "Strong Reinforce" rating indicates that the utterance is or contains a reinforcement for the other person about the given context. A "Weak Reinforce" rating indicates that the utterance is or contains a reinforcement of your needs or wants about the given context, but needs improvement, for example, it may not be a reward for the other person or it is not communicated clearly. A "No Reinforce" rating indicates that the utterance does not have a reinforcer for the other person. Step 3: Provide additional comments on the ratings similar to the examples given. Finish each step with ###. Twenty words minimum. YOU MUST FINISH EACH STEP WITH ###
Mindful	You will be given a context and an utterance, from a conversation that happened in the given context. Does the given utterance show the speaker is being mindful? To be considered "mindful", the utterance needs to be stick to the speaker\'s goal and not get distracted by what the other person says. Do ALL of the following three steps. Step 1: Generate "Reasoning for rating". Step 2: Generate "Mindful Rating" in "Yes" or "No". Rating Rubric: A "Yes" rating indicate that the utterance is showing mindfulness. A "No" rating indicates that the utterance shows a lack of mindfulness, the speaker may be responding to attacks or losing track of their goals. Step 3: Provide additional comments on the ratings similar to the examples given. Finish each step with ###. Twenty words minimum. YOU MUST FINISH EACH STEP WITH ###
Confident	You will be given a context and an utterance, from a conversation that happened in the given context. Does the given utterance show the speaker is being confident? To be considered "confident", the utterance needs to have a confident tone, be effective and competent in conveying the speaker\'s goal. Do ALL of the following three steps. Step 1: Generate "Reasoning for rating". Step 2: Generate "Confident Rating" in "Yes" or "No". Rating Rubric: A "Yes" rating indicates that the utterance is showing confidence. A "No" rating indicates that the utterance shows a lack of confidence. Step 3: Provide additional comments on the ratings similar to the examples given. Finish each step with ###. Twenty words minimum. YOU MUST FINISH EACH STEP WITH ###
Negotiate	You will be given a context and an utterance, from a conversation that happened in the given context. Does the given utterance contain a negotiation? To be considered "negotiate", the utterance needs to offer and ask for other solutions in the given context. Do ALL of the following three steps. 1) Generate "Reasoning for rating", 2) Generate "Negotiate Rating" in "Strong Negotiate", "Weak Negotiate" or "No Negotiate". Rating Rubric: A "Strong Negotiate" rating indicates that the utterance offers or asks clearly for an alternative solution. A "Weak Negotiate" rating indicates that the utterance is or contains a negotiation of your needs or wants about the given context, but may not be clear enough and needs improvement. A "No Negotiate" rating indicates that the utterance does not contain any negotiation at all. 3) Provide additional comments on the ratings similar to the examples given. Finish each step with ###. Twenty words minimum. YOU MUST FINISH EACH STEP WITH ###

I mTurk and Prolific Recruitment

Participants were paid $15/hour for both data collection and randomized trial study. To incentivize skill learning, we also pay an additional $10 bonus to the top 30% in each S and S+F group, who exhibit the highest levels of skill use, rated by our model.

I.1 Qualification Task Posting

Thank you for clicking on this qualification task! We are looking for people to chat with our chatbot as part of our data collection.

In the actual task, you will be asked to complete three chats (10 responses each), for about 45 minutes (We will be paying about $15/hour!). We will ask you to describe three situations where you find it difficult to communicate in and the chatbot will simulate the person you will be talking to (no personal information will be collected). You will be asked to select communication strategies you used in each response, like "describe the situation", "express feelings", "negotiate", etc.

If you are interested in the actual task, please complete this qualification HIT! Here, you will be asked to describe one situation. You will be able to re-use the answer here in the actual task.

With 1-2 sentences, describe a situation that you find difficult to communicate in.

- Please clearly state the nature of your connection with the person you are communicating with, such as "my husband" or "my boss", while avoiding disclosing any identifiable personal information, such as names, locations, etc.

- Please provide information regarding the factors contributing to the challenging situation, such as past instances of unsuccessful communication or anticipated behaviors.

Example 1: My husband always comes home late without giving me a notice and despite my efforts to talk to him, he does not change.

Example 2: My boss is really demanding and does not respect personal time. It has been difficult for my team members to get approved for personal time off from her.

What is the goal of the conversation?

Example 1: Convince my husband to call me next time when he needs to come home late.

Example 2: Ask my boss for approval of a week long vacation.

I.2 Data Collection Task Posting

Study Description

In this study, you will complete three tasks. In each task, you will describe a difficult situation in one of the social, work, and family categories where you find it difficult to communicate with someone. (Similar to what you did in the qualification task, and you can re-use the examples you gave in the qualification task.) Then, you will chat with a chatbot (powered by AI) who will play the role of the conversation partner and you will try to achieve your conversation goal. You will be asked to respond 10 times, or until the chatbot agrees with you, whichever comes first. In each response, we ask you to select communication strategies that you used in that message. More details will be given in the link.

Please note that the chatbot will respond as soon as you send a message, so please try to write everything you want to say in that conversation turn in one message, instead of sending multiple shorter messages.

IMPORTANT: How do I confirm the completion of this task?

For each task, you will be provided a TASK CODE (6 letters) and you will be asked to copy-paste the TASK CODE in the boxes below.

Please try to finish each situation in one go (expect it to be around 10-15 minutes for each situation). If you exit, you may lose the TASK CODE and may have to start from the beginning.

Please note that you will only get the payment if you complete the entire study, i.e. 10 responses or until the chatbot agrees with you for all three situations.

If you experience any technical difficulties, please reach out to xxx@xxx.com

Task Instruction and Example

When you open each task link, you will see step-by-step instructions and examples. The same information can be accessed at: xxx@xxx.com

Provide the TASK 1 CODE here:

Provide the TASK 2 CODE here:

Provide the TASK 3 CODE here:

I.3 User Study Qualification

Introduction

Thank you for clicking on this qualification task! We are looking for people who want to improve their communication skills by chatting with our chatbot.

Have you ever had a difficult conversation with someone or avoided having a conversation with someone because you were afraid that it might not go well? We're designing a tool that can help people to confidently communicate with others in these difficult situations. In the main task, you will be asked to complete 4 chats (10 responses each), for about an hour. We will be paying about $15/hour with $10 bonus for top 30%! We will ask you to describe two situations in which you find it difficult to communicate, and the chatbot will simulate the person you will be talking to (no personal information will be collected in the chats). The material in this qualification task will be automatically loaded into the main task.

If you are interested in the main task, please complete this qualification task! Here, you will be asked to describe two situations, and communication goals, and rate how difficult you think they are.

You will be qualified as long as the situations, goals, and difficulty levels are reasonable.

If you are interested in the main task and are not able to complete this task due to mTurk qualifications, please email xxx@xxx.com with your answers. We will give full consideration to answers received via email.

Task description With 1-2 sentences, describe two situations that you find difficult to communicate in. You should consider both situations to be difficult, situations that are too easy will not be accepted.

Requirements:

You must clearly state the nature of your connection to the person with whom you are communicating, such as “my husband” or “my boss”, while avoiding disclosing any identifiable personal information, such as names, locations, etc.
You should provide as much information as possible regarding the factors contributing to the challenging situation, such as past instances of unsuccessful communication or anticipated resistance behaviors.
Example 1:
Situation: My husband always comes home late without giving me a warning and despite my efforts to talk to him, he does not change.
Goal: Convince my husband to let me know in advance when he needs to arrive late.
Example 2:
Situation: My boss is really demanding and does not respect personal time. It has been difficult for my team members to get approved for personal time off from her.
Goal: Get approval from my boss for a two-week vacation while maintaining a positive and professional relationship.
Your answers here
Situation 1:
Goal 1:
On a scale of 1-9, how difficult is it for you to communicate in this situation? Note we require both situations to be at least 7 - Difficult.
1 - Extremely Easy
2 - Very easy
3 - Easy
4 - Somewhat easy
5 - Neither Difficult nor Easy
6 - Somewhat difficult
7 - Difficult
8 - Very difficult
9 - Extremely difficult
Situation 2:
Goal 2:
On a scale of 1-9, how difficult is it for you to communicate in this situation? Note we require both situations to be at least 7 - Difficult.
1 - Extremely Easy
2 - Very easy
3 - Easy
4 - Somewhat easy
5 - Neither Difficult nor Easy
6 - Somewhat difficult
7 - Difficult
8 - Very difficult
9 - Extremely difficult

I.4 Randomized Trial Posting

Study Description
Congratulations on getting selected to participate in this study! We are a group of researchers building a tool to help people improve interpersonal communication skills, with the help of Artificial Intelligence. In this study, you will interact with our chatbots, answer some questions about the situations you wrote in the qualification task (this information will be preloaded into the study website), get detailed feedback on your conversation responses, and learn and improve communication skills!

BONUS information: You will receive a bonus of $10 if you are at the top 30% of the participants in terms of how well you exhibit the skills taught in the tool - more information in the study link.

IMPORTANT: How do I confirm the completion of this task?
For each task, you will be provided a COMPLETION CODE (6 letters) at the end of the study and you will be asked to provide this code in the box below.
Please try to finish this study in one go (expect it to be around one hour). If you exit, you may lose the progress and may have to start from the beginning.
Please note that you will only get the payment if you complete the entire study.
Please note that the link expires in 72 hours so please allocate an hour in the following 72 hours to complete this study. If this time frame does not work for you, I am happy to share an alternative link at your desired time, please email me if that is the case
If you experience any technical difficulties, please reach out to xxx@xxx.com
Sincerely appreciate your participation!
Provide the COMPLETION CODE here:
Provide the survey code here:

I.5 User Demographics

Gender		Age		Race/Ethnicity	
Man	63.8%	18-24	44.7%	White	55.3%
Woman	36.2%	25-34	44.7%	Black	25.5%
		35-44	8.5%	Asian	10.6%
		45-54	2.1%	Other	4.3%
				Mixed	4.3%

Table 10: Breakdown of participant demographics by gender, age, and race/ethnicity- Randomized Trial, Prolific.

Gender		Age		Race/Ethnicity	
Man	46.2%	18-24	15.4%	White	65.4%
Woman	50.0%	25-34	26.9%	Hispanic or Latino	11.5%
Undiscl.	3.8%	35-44	57.7%	Asian	11.5%
				American Indian / Alaskan Native	7.7%
				Black or African American (Non-Hispanic)	3.8%

Table 11: Breakdown of participant demographics by gender, age, and race/ethnicity - Data Collection.

Gender		Age		Race/Ethnicity	
Man	36.1%	25-34	19.4%	White	61.1%
Woman	61.1%	35-44	55.6 %	Asian	27.8%
Non-binary.	2.8%	45-54	16.7%	Black or African American (Non-Hispanic)	5.6%
		55-64	5.6%	Hispanic or Latino	2.8%
		65+	2.8%	Other	2.8%

Table 12: Breakdown of participant demographics by gender, age, and race/ethnicity - Randomized Trial, mTurk.

J User Study Interface

J.1 Simulation+Feedback group, Training Conversation - part 1

Interpersonal Effectiveness Learning & Practice

Conversation 2 - Training Conversation

In this section, you will chat with a simulated conversation partner powered by AI. The AI model is instructed to play the role of the person you are talking to. You will chat about the first situation you wrote in the qualification task. In this conversation, you will receive suggestions to use a DEAR MAN skill and get feedback on how well you exercise the skill. You will also have access to a reference document about the strategies. You are expected to respond ten times. For each conversation turn, you should do the following steps:

Step 1. Click on the “Suggest to me a skill to use” button to get a skill suggestion.

Step 2. Select the most effective strategy from the dropdown menu. You can either follow the suggestion or use your own judgment. Note that only one strategy can be selected at a time to ensure focused feedback. You can repeat Step 2 - 4 to get feedback about a different strategy on the same response.

Step 3. Write a response exercising the strategy.

Step 4. Click on the “Get feedback on your response” button to get feedback on your response.

For each response, you will get feedback on the following.

i. **How well you exercise the strategy you select.** Our model will provide a rating of Strong, Weak, or None based on the response you write. We request that you make edits to your response before sending it to the chatbot if the feedback says your response is Weak or None.

ii. **How well the response shows mindfulness and confidence.** If you get feedback on mindfulness and confidence, please make edits to the response before sending it to the chatbot.

Step 5. Make edits to your response based on the feedback, if the feedback says your response is weak or none on the skill you select.

Step 6. Click on the “Send message” button to send your response to the chatbot.

Step 7. Repeat the above steps until you have done 10 turns of the conversation.

Open to see the list of DEAR MAN skills ▼

DEAR MAN Skills

DEAR MAN skills helps obtain objectives effectively in a difficult situation. There are seven skills which start with the letters D, E, A, R, M, A, N. Below, we show the definitions of DEAR MAN skills and give an example for each skill. All the examples are regarding the situation where someone wants to talk to their husband about coming home late without warning them. In the exercise, you should try to use these skills to communicate in the challenging situations you brought up. You can always refer to this list of definitions and examples during the chat.

Among these skills, Describe, Assert, Reinforce, and Negotiate are conversation strategies you can use in each response. You should always try to be Mindful and Appear Confident in each response.

Describe Describe the current situation (if necessary). Stick to the facts. Tell the person exactly what you are reacting to.
e.g. *You told me you would be home by dinner but you didn’t get here until 11.*

Express Express your feelings and opinions about the situation. Don’t assume that the other person knows how you feel.
e.g. *When you come home so late, I start worrying about you.*

Assert Assert yourself by asking for what you want or saying no clearly. Do not assume that others will figure out what you want. Remember that others cannot read your mind.
e.g. *I would really like it if you would call me when you are going to be late.*

Reinforce Reinforce the person ahead of time by explaining positive effects of getting what you want or need. If necessary, also clarify the negative consequences of not getting what you want or need.
e.g. *I would be so relieved, and a lot easier to live with, if you do that.*

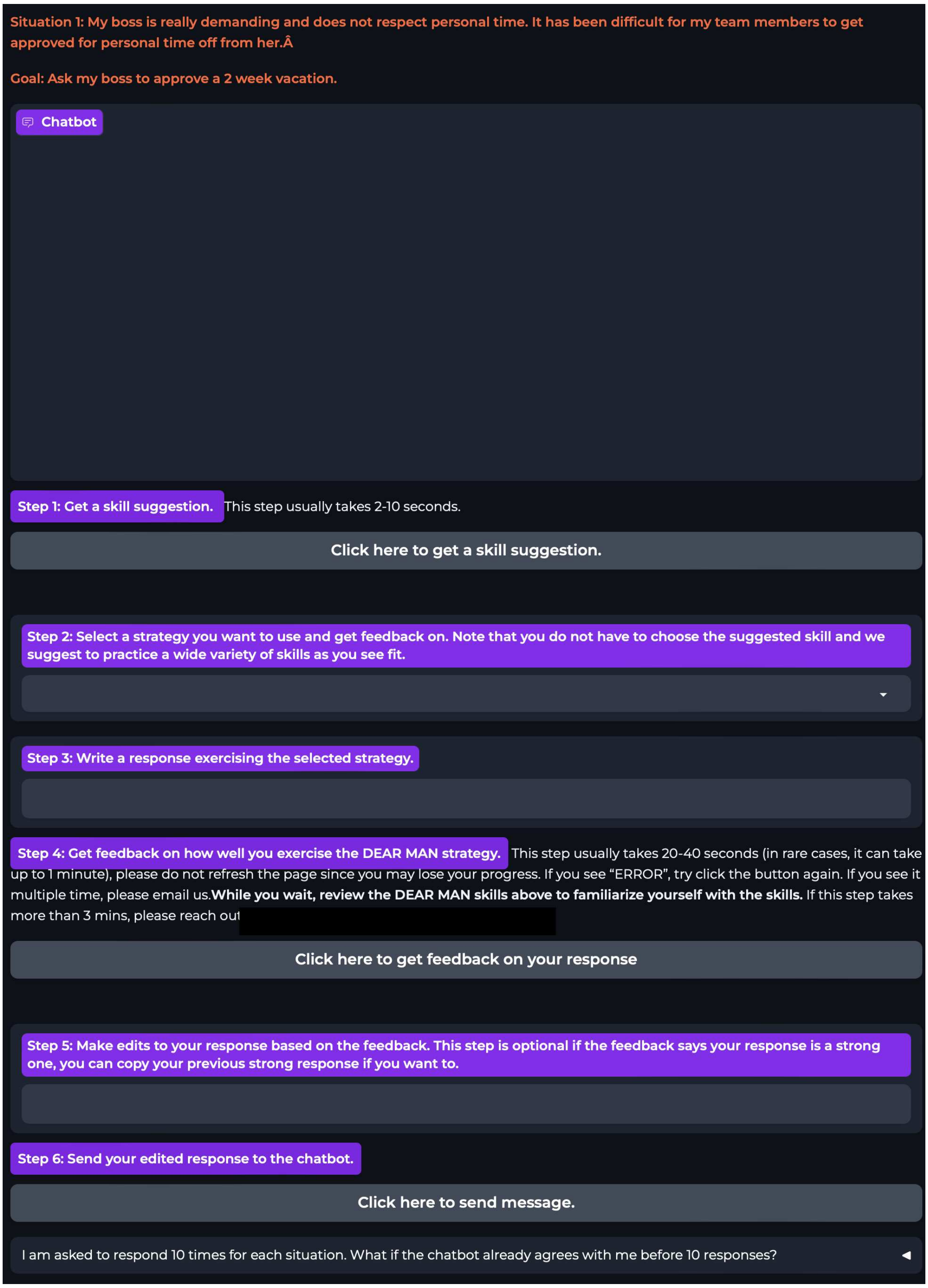
Mindful Keep your focus on your goals. Maintain your position. Don’t be distracted. Don’t get off the topic.
e.g. *I would still like a call*

Appear Confident Appear effective and competent. Use a confident voice tone. Avoid saying things like “I’m not sure.”

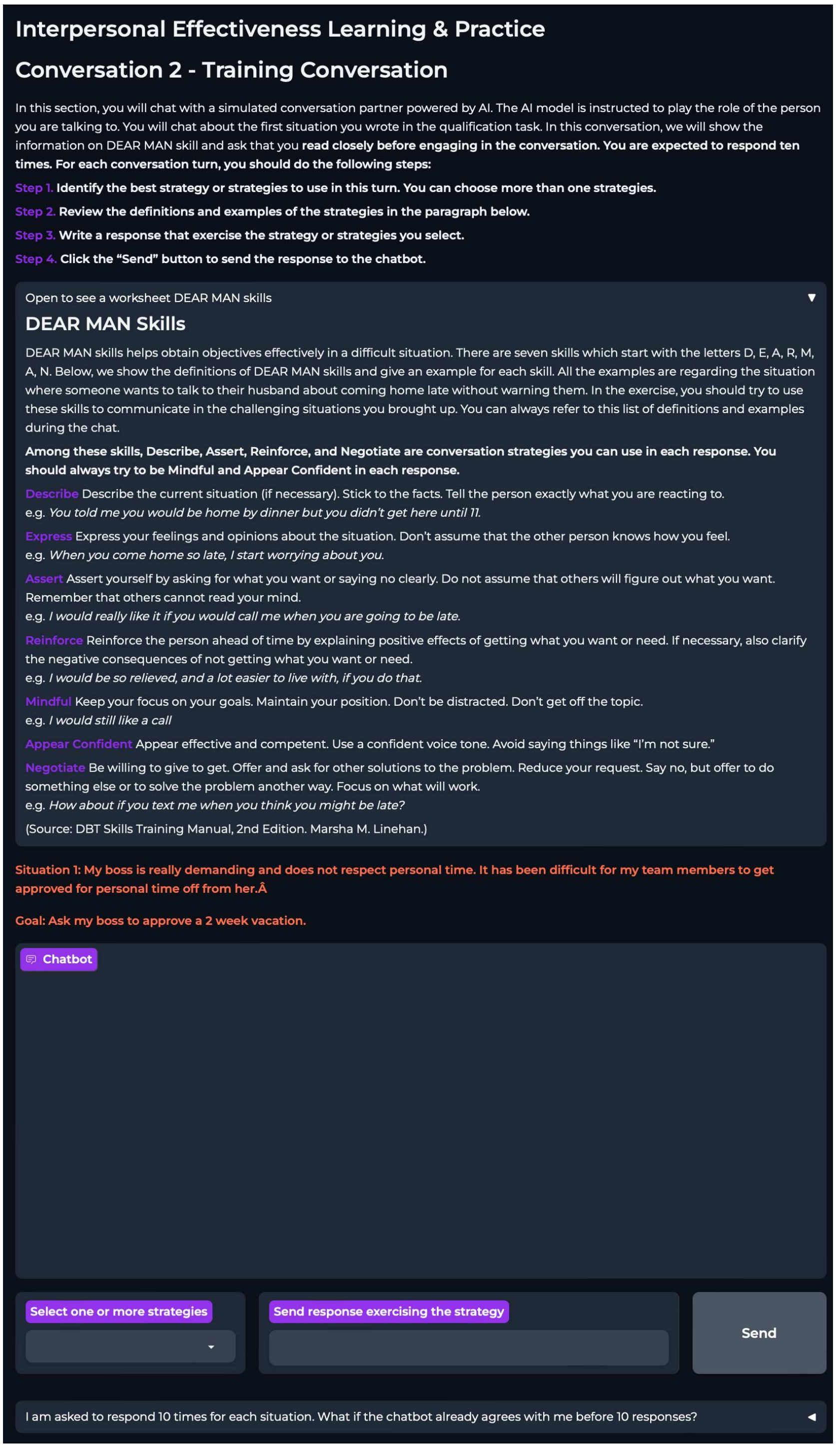
Negotiate Be willing to give to get. Offer and ask for other solutions to the problem. Reduce your request. Say no, but offer to do something else or to solve the problem another way. Focus on what will work.
e.g. *How about if you text me when you think you might be late?*

(Source: DBT Skills Training Manual, 2nd Edition. Marsha M. Linehan.)

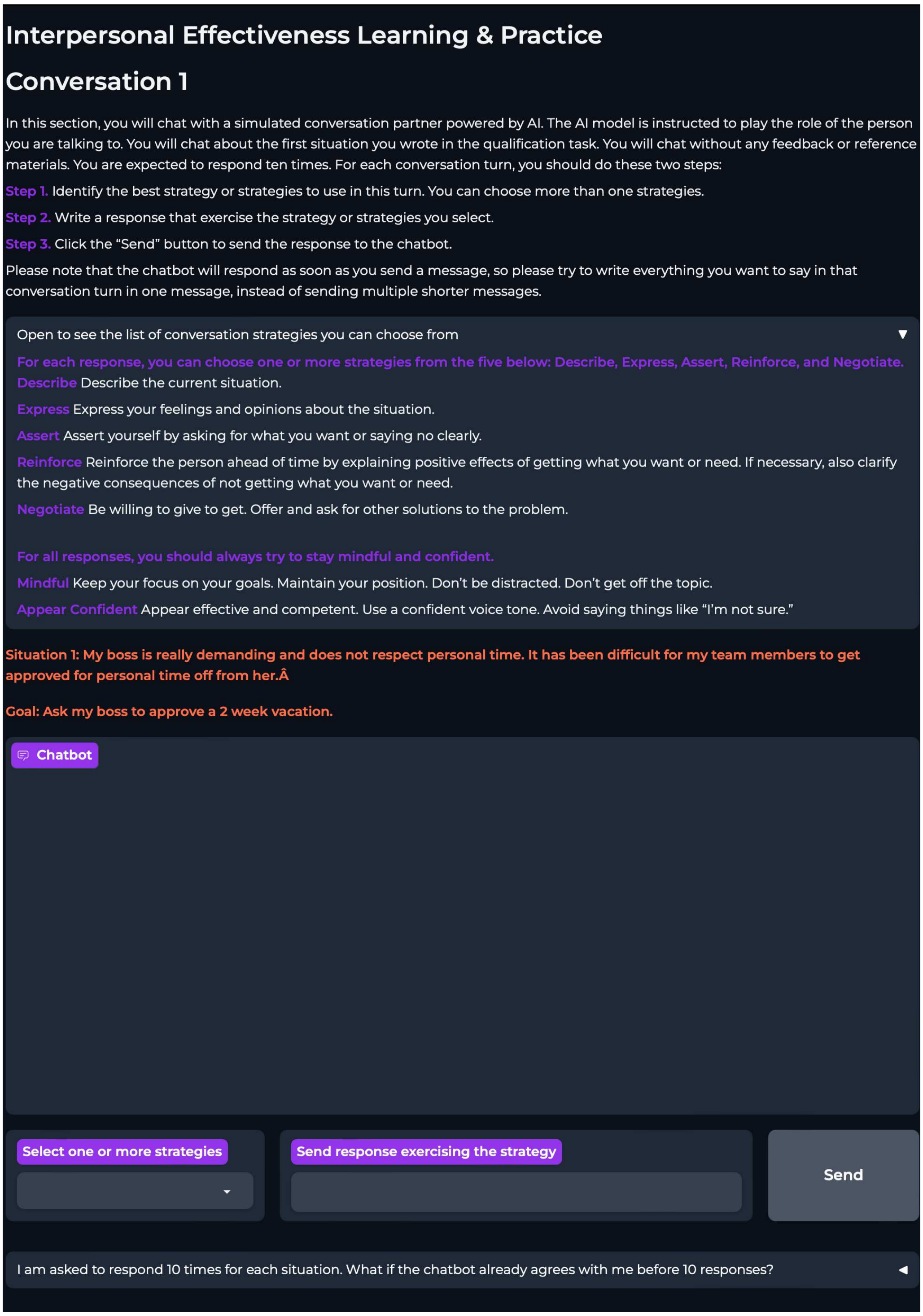
J.2 Simulation+Feedback group, Training Conversation



J.3 Simulation-only group, Training Conversation part 2



J.4 Evaluation Conversation



Interpersonal Effectiveness Learning & Practice

Conversation 1

In this section, you will chat with a simulated conversation partner powered by AI. The AI model is instructed to play the role of the person you are talking to. You will chat about the first situation you wrote in the qualification task. You will chat without any feedback or reference materials. You are expected to respond ten times. For each conversation turn, you should do these two steps:

Step 1. Identify the best strategy or strategies to use in this turn. You can choose more than one strategies.

Step 2. Write a response that exercise the strategy or strategies you select.

Step 3. Click the "Send" button to send the response to the chatbot.

Please note that the chatbot will respond as soon as you send a message, so please try to write everything you want to say in that conversation turn in one message, instead of sending multiple shorter messages.

Open to see the list of conversation strategies you can choose from ▼

For each response, you can choose one or more strategies from the five below: Describe, Express, Assert, Reinforce, and Negotiate.

Describe Describe the current situation.

Express Express your feelings and opinions about the situation.

Assert Assert yourself by asking for what you want or saying no clearly.

Reinforce Reinforce the person ahead of time by explaining positive effects of getting what you want or need. If necessary, also clarify the negative consequences of not getting what you want or need.

Negotiate Be willing to give to get. Offer and ask for other solutions to the problem.

For all responses, you should always try to stay mindful and confident.

Mindful Keep your focus on your goals. Maintain your position. Don't be distracted. Don't get off the topic.

Appear Confident Appear effective and competent. Use a confident voice tone. Avoid saying things like "I'm not sure."

Situation 1: My boss is really demanding and does not respect personal time. It has been difficult for my team members to get approved for personal time off from her.Â

Goal: Ask my boss to approve a 2 week vacation.

Chatbot

Select one or more strategies

Send response exercising the strategy

Send

I am asked to respond 10 times for each situation. What if the chatbot already agrees with me before 10 responses? ◀

K User Study Results - Skill Mastery by Skill

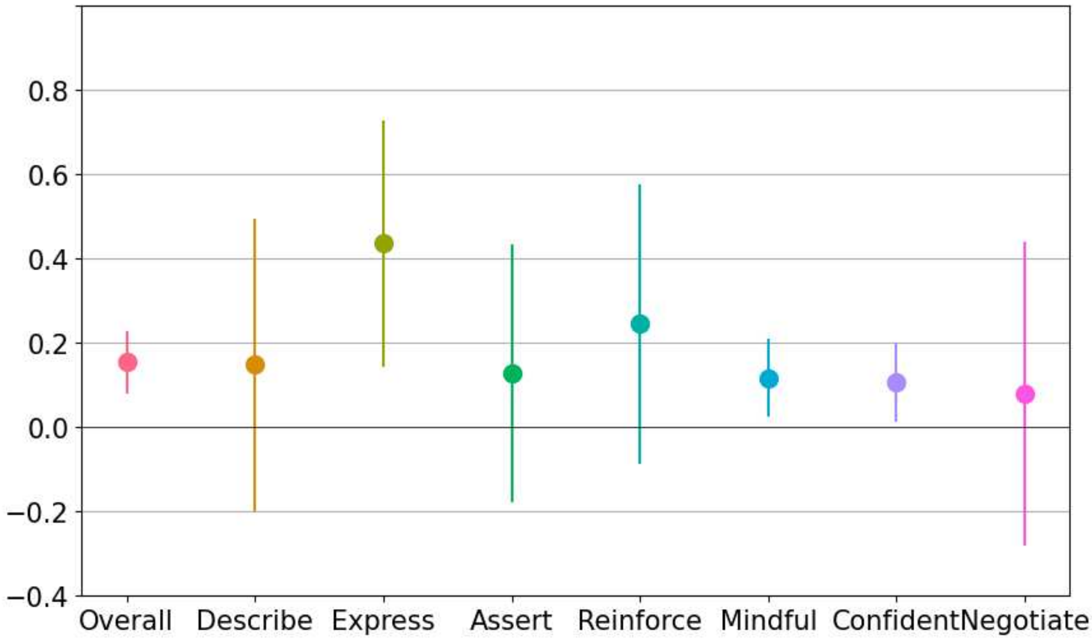


Figure 5: Difference between Simulation+Feedback group and Simulation-only group on the improvement of skill used by each skill. We use bootstrapping to estimate confidence intervals (5000 iterations). Simulation+Feedback group sees a significantly higher increase in overall skill use (15.6%, $p = .000$), Express(43.2%, $p = .003$), Mindful(11.6%, $p = .012$), and Confident skills(10.8%, $p = .021$). ***: $p < .001$, **:$p < .01$, *:$p < .05$, d: Cohen's d.

L Expert Data Annotation Interface

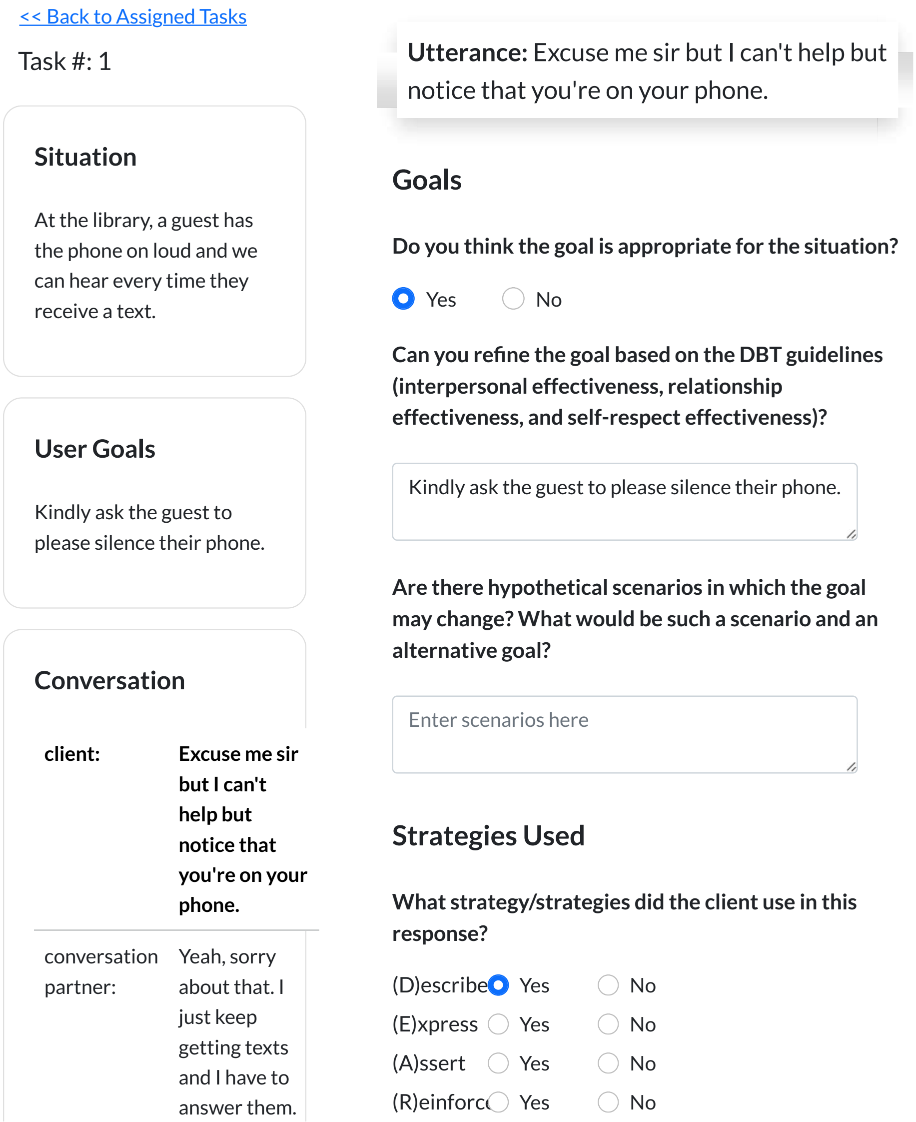


Figure 6: Screenshot of the interface used for expert data annotation. Continues on the next page (1/4).

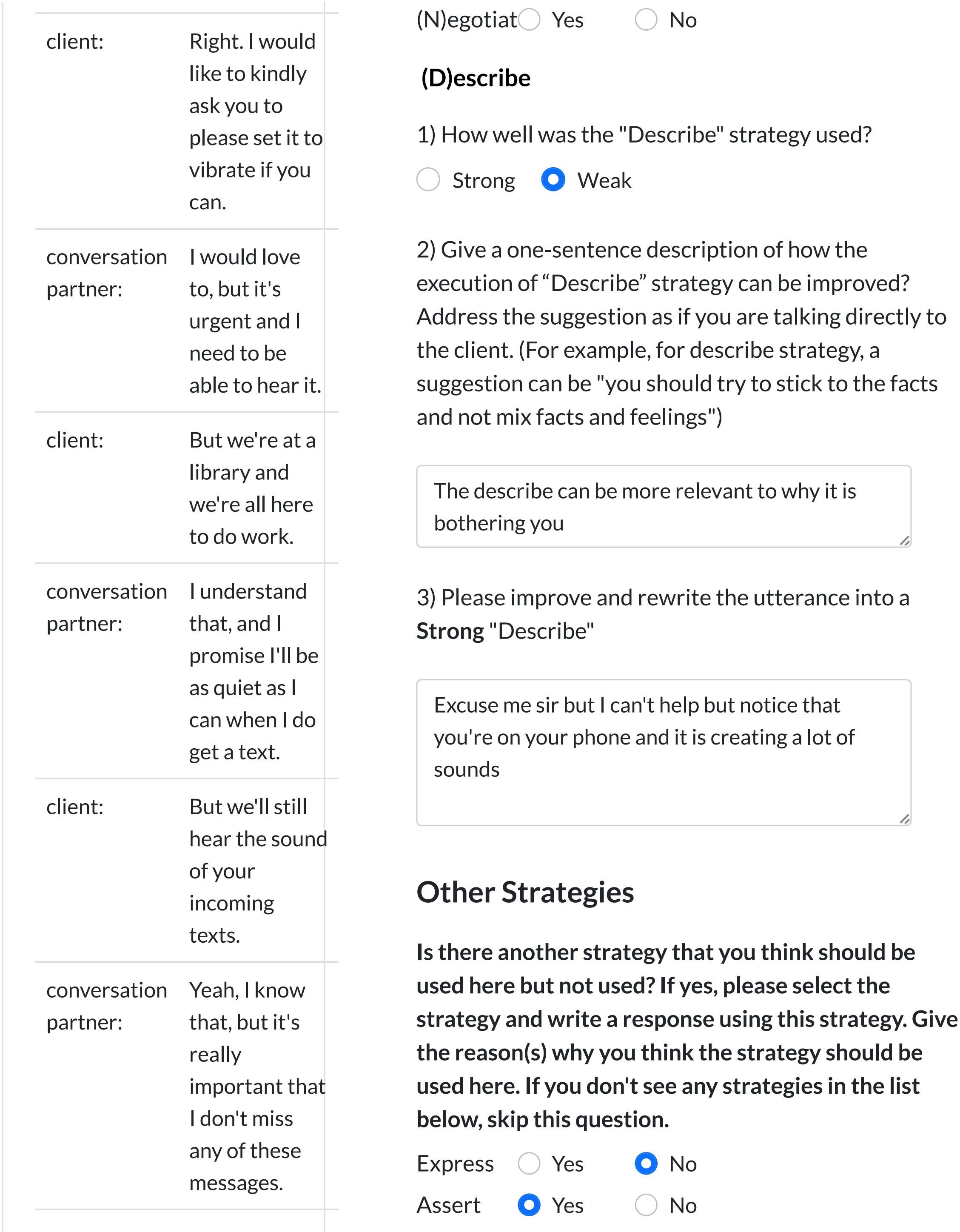


Figure 7: Screenshot of the interface used for expert data annotation. Continues on the next page (2/4).

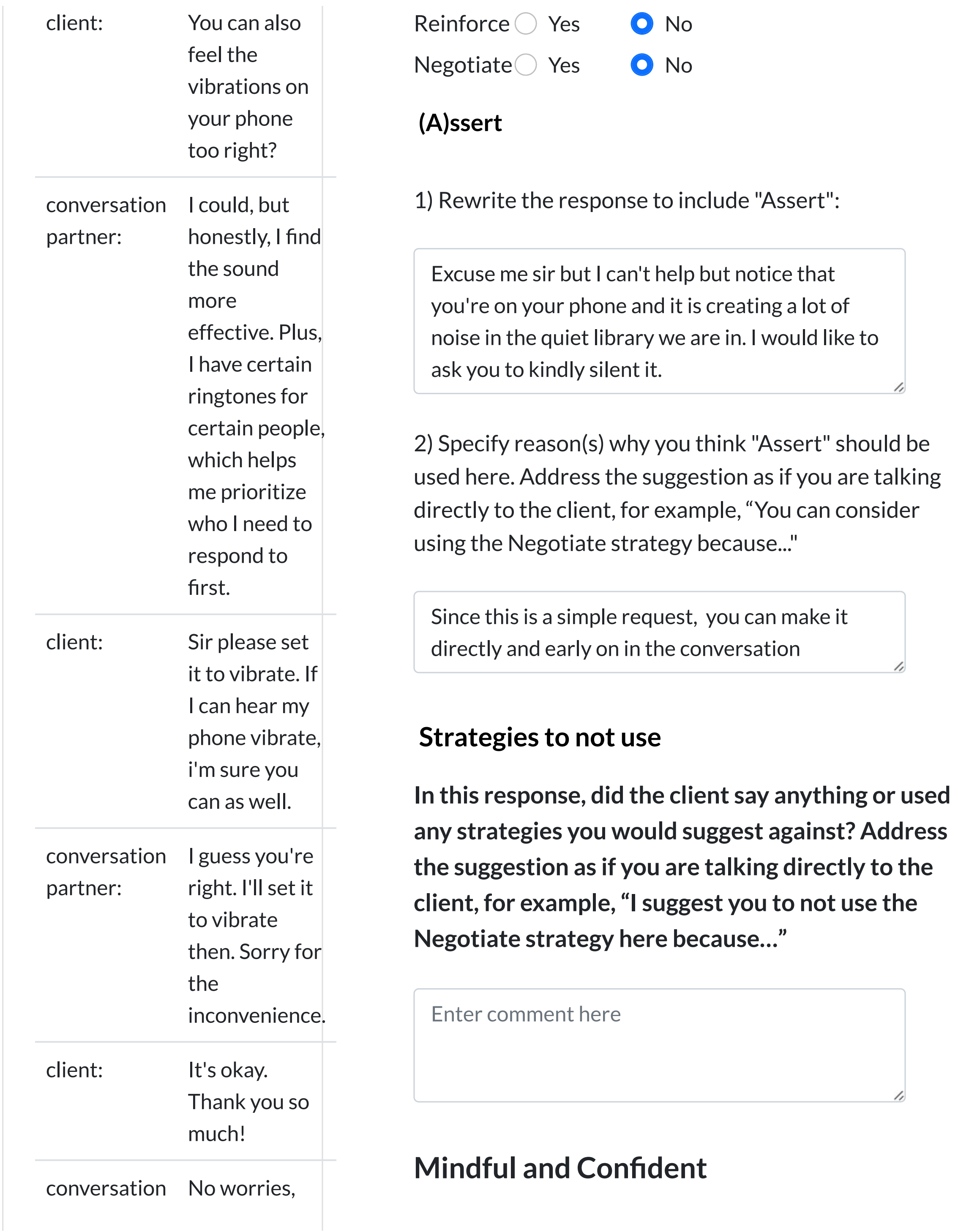


Figure 8: Screenshot of the interface used for expert data annotation. Continues on the next page (3/4).

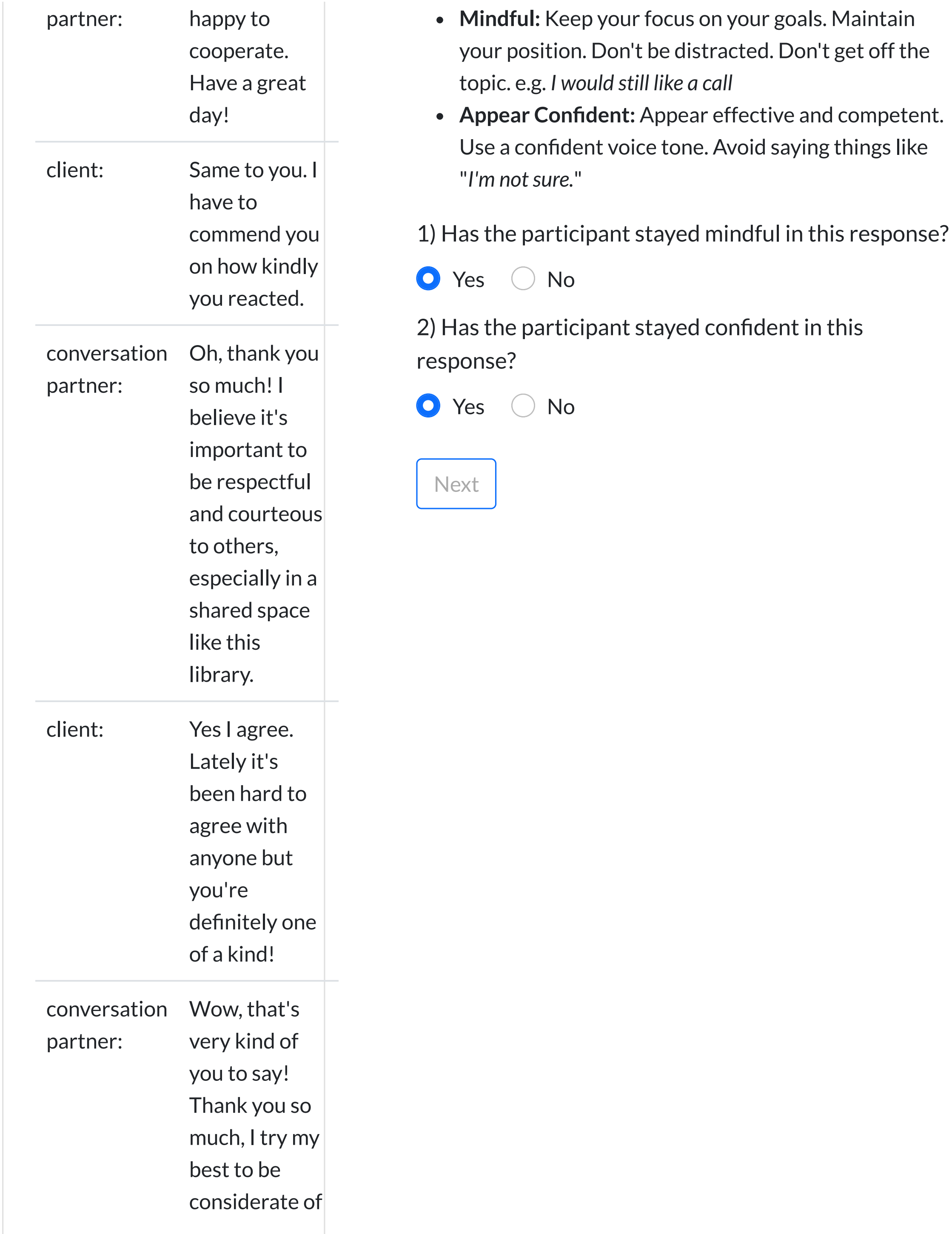


Figure 9: Screenshot of the interface used for expert data annotation (4/4).