Multi-Level Feedback Generation with Large Language Models for Empowering Novice Peer Counselors

Alicja Chaszczewicz [S], **Raj Sanjay Shah*** [GT], **Ryan Louie*** [S]
Bruce A Arnow [S], **Robert Kraut** [CMU], **Diyi Yang** [S]
Stanford University [S], Georgia Institute of Technology [GT], Carnegie Mellon University [CMU]

Abstract

Realistic practice and tailored feedback are key processes for training peer counselors with clinical skills. However, existing mechanisms of providing feedback largely rely on human supervision. Peer counselors often lack mechanisms to receive detailed feedback from experienced mentors, making it difficult for them to support the large number of people with mental health issues who use peer counseling. Our work aims to leverage large language models to provide contextualized and multi-level feedback to empower peer counselors, especially novices, at scale. To achieve this, we co-design with a group of senior psychotherapy supervisors to develop a multi-level feedback taxonomy, and then construct a publicly available dataset with comprehensive feedback annotations of 400 emotional support conversations. We further design a self-improvement method on top of large language models to enhance the automatic generation of feedback. Via qualitative and quantitative evaluation with domain experts, we demonstrate that our method minimizes the risk of potentially harmful and low-quality feedback generation which is desirable in such high-stakes scenarios.

1 Introduction

Realistic practice and tailored feedback are key processes for training peer counselors with clinical skills. Providing feedback could significantly enhance peer counselor skills, thereby improving support quality and benefiting many seeking help online (Ali et al., 2015). However, it is often time-consuming and costly for counseling supervisors to provide detailed feedback (Atkins et al., 2014) to beginner peer counselors. Without appropriate guidance, peer counselors might develop biased or even inappropriate counseling helping skills without being aware of it, based on their own experiences. What can we do to provide detailed feedback to a large number of novice peer counselors at scale? In this work, we explore whether large language models (LLMs) can be used to provide contextualized feedback to empower peer counselors in training.

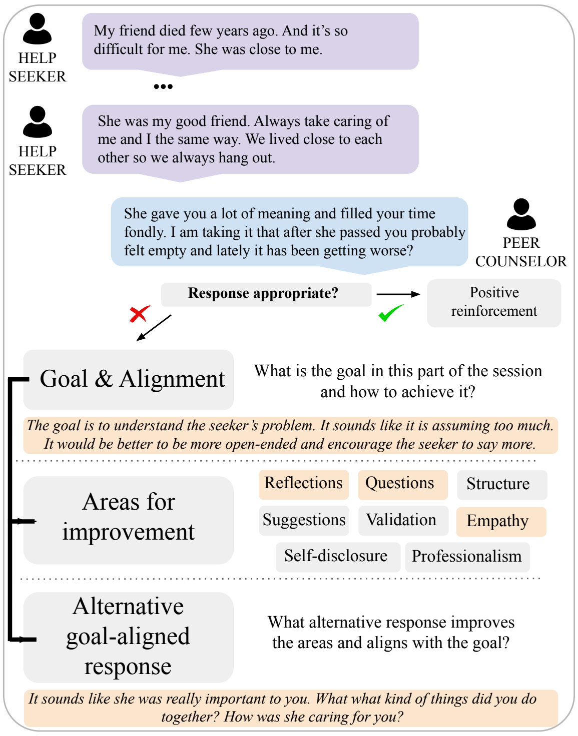


Figure 1: Example conversation excerpt taken from the ESConv dataset (Liu et al., 2021) annotated using our feedback taxonomy. Feedback components (*appropriateness, goal definition and alignment, areas for improvement, alternative goal-aligned response*) are demonstrated on one utterance of the peer counselor's response (in blue). Optionally, one can also provide *positive reinforcement* by highlighting areas in categories peer counselors excelled at.

Numerous recent studies have explored the feasibility of applying computational techniques to differentiate between low and high-quality counseling

Email: {alicjach, rylouie, arnow, diyiy}@stanford.edu, rajsanjayshah@gatech.edu, robert.kraut@cmu.edu
* These authors contributed equally to this work

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4130–4161
August 11-16, 2024

automatically (Pérez-Rosas et al., 2019; Imel et al., 2019; Sharma et al., 2020; Flemotomos et al., 2021; Min et al., 2022; Shen et al., 2022; Wu et al., 2023; Fang et al., 2023; Sharma et al., 2023; Hsu et al., 2023; Chiu et al., 2024). In doing so, prior work mostly provides numeric feedback to counselors about how well a particular skill is used. Some recent studies provide utterance-level suggestions of responses to use according to appropriate counseling helping skills (Hsu et al., 2023), or alternatives for more empathetic responses (Sharma et al., 2023). Yet, little attention is given to developing automatic feedback that closely mirrors how clinical supervisors provide feedback to novice counselors.

To this end, we co-designed a feedback framework with senior psychotherapy supervisors to reflect the content and delivery of feedback they give to novice counselors. Concretely, we conducted a contextual inquiry (Karen and Sandra, 2017) with supervisors engaging in a representative task of providing feedback on a transcript of an emotional support conversation (Liu et al., 2021) as if they were communicating the feedback to a novice counselor. We then developed a multi-level feedback framework by modeling the common patterns at different granularity observed in interviews and important feedback dimensions highlighted in textbooks and training for foundational active listening skills (Hill, 2009; 7Cups, 2023). With this multi-level feedback framework presented in Figure 1, we introduce a publicly available dataset of conversations enriched with comprehensive feedback annotations, building upon an existing public emotional support conversations dataset ESConv (Liu et al., 2021). Specifically, we leverage a model-in-the-loop annotation paradigm where GPT-4 and counseling domain experts work together to produce the annotations for 400 conversations.

To enable transparent model development, especially for a high-stakes domain like counseling, we fine-tuned the open-source Llama-2 model to generate multi-level feedback. We further introduce a simple but effective self-improvement method to forecast how specific feedback might improve subsequent interaction and use this forecast information to supervise feedback generation. Unlike general natural language generation tasks, we aim at optimizing feedback generation for worst-case performance since failures (e.g., generating poor advice) matter more in this high-stakes scenario. Using both quantitative evaluation and qualitative evaluation with domain experts, we demonstrate that our approach generates high-quality feedback and significantly boosts the worst-case performance on multi-level feedback generation compared to baselines. In summary, this paper makes the following contributions:

- We propose a novel and comprehensive multi-level feedback framework for training peer counseling skills co-designed with senior psychotherapy supervisors.
- We constructed and make publicly available *FeedbackESConv*[1], a dataset of 400 emotional support conversations with multi-level feedback annotated by domain experts and GPT-4.
- We enhanced a fine-tuned LLM for multi-level feedback using a simple but effective self-improvement method to forecast how specific feedback might improve subsequent interaction and further use such signals to supervise the feedback generation.
- Via extensive evaluations with domain experts, we found that, compared to baselines, it significantly boosts the worst-case performance on multi-level feedback generation.

2 Related Work

There have been different approaches to building automated methods that help peer counselors improve their skills, ranging from scoring-based methods (e.g., measures of empathy; the use of counseling-specific dialogue acts) to automatically generated suggestions for alternative responses.

Scoring for assessment (Pérez-Rosas et al., 2019) train a classifier to distinguish between high-quality and low-quality YouTube and Vimeo Motivational Interviewing videos. (Tanana et al., 2019) introduce a system that detects reflections and open questions used by a trainee in a simulated chat and scores the whole conversation by proving the percentage frequency of those. (Imel et al., 2019) combine technical methods of (Xiao et al., 2015; Tanana et al., 2016, 2019) and develop a system for numerical scoring of Motivational Interviewing skills, amount of empathy, and number of reflections and open questions. (Sharma et al., 2020) design a model to identify and quantify the level

[1]We provide our code at https://github.com/SALT-NLP/counseling-feedback

	Numerical scoring of response quality	Suggestion of response or alternate response	Response evaluation across multiple peer counseling skills categories	Goal - oriented natural language explanations
Pérez-Rosas et al. (2019)	✓	✗	✓	✗
Tanana et al. (2019); Imel et al. (2019)	✓	✗	✓	✗
Sharma et al. (2020)	✓	✗	✗	✗
Flemotomos et al. (2021)	✓	✗	✓	✗
Min et al. (2022)	✓	✗	✗	✗
Shen et al. (2022)	✗	✓	✗	✗
Sharma et al. (2023)	✗	✓	✗	✗
Min et al. (2023); Welivita and Pu (2023)	✗	✓	✗	✗
Hsu et al. (2023)	✗	✓	✓	✗
Chiu et al. (2024)*	✓	(✓)	✓	✗
Our work	✓	✓	✓	✓

Table 1: Categorization of previously proposed approaches aimed at evaluating or enhancing the quality of emotional support conversations. "Numerical scoring of response quality" indicates whether a study applied a binary or continuous scale for quality assessment. "Response evaluation across multiple peer counseling skills categories" indicates whether the feedback mechanism incorporated a multidimensional structure (more than two dimensions). "Suggestion of response" examines if the approach includes generating potential peer counselor answers. "Goal-oriented natural language explanation" indicates whether the system offers natural language conversation goals and explains how errors it identified can be aligned to these goals. *Chiu et al. (2024) is concurrent work focusing on evaluating the quality of LLM-based therapy simulations.

of empathy in emotional support conversations and (Flemotomos et al., 2021) develop a system to assess the quality of real therapy sessions by automatically rating transcripts using the Cognitive Therapy Rating Scale. (Min et al., 2022) focus on Motivational Interviewing reflection skills and train a model that assigns a score in the range (0,1) to evaluate the reflection quality. In concurrent work to ours (Chiu et al., 2024) use GPT-4 with in-context learning to solve the task of multi-classification of conversational behaviors of clients and therapists.

Generation of suggestions (Shen et al., 2022) create a model to automatically generate good reflections. (Min et al., 2023) introduce a system that rewrites non-reflective counseling responses to transform them into reflective ones. (Welivita and Pu, 2023) build a model that rephrases counseling responses that contain advice without permission into ones that adhere to Motivational Interviewing guidelines. (Sharma et al., 2023) develop HAILEY AI-agent that provides just-in-time suggestions on how to respond more empathetically. (Hsu et al., 2023) introduce the CARE system that generates response suggestions based on predicted appropriate Motivational Interviewing strategies.

Rather than taking a technical perspective focusing on the feedback systems that *can* be built with scoring or response generation methods, we posit that one can design better-automated feedback methods for peer counseling training by understanding and mirroring the existing ways supervisors deliver feedback to novices. Our co-design reveals that post-session feedback for peer counseling *encompasses and extends beyond* scoring and suggestions for improving individual response quality. Crucially, it emphasizes that each response should be based on the counseling goals it should serve at the specific point in the session. Incorporating contextualized *goals* into the feedback structure provides a purpose-led orientation (see Table 1). Such natural language goal descriptions are especially valuable since providing explanations is more beneficial for learning than simply giving the correct answer (Butler et al., 2013).

2.1 Generation Capabilities of LLMs

Past work explored the capabilities of LLMs in generating natural-language feedback across various domains. Wang et al. (2023a) explores the use of LLMs like GPT-4 and GPT3.5 in math tutoring to deliver high-quality feedback to remediate student mistakes. Liang et al. (2023) employ GPT-4 for generating comprehensive reviews for research papers. These varied applications demonstrate the adaptability and potential of LLMs to generate feedback across educational and professional settings. Unlike past work that builds feedback systems directly on top of GPT-4, we seek to enable the transparent development of open-source feedback models for the domain of peer counseling.

Thus, we first develop an annotated dataset of feedback which is co-annotated by domain experts and GPT-4 using our multi-level feedback taxonomy, and then fine-tune the open-source Llama2-13B model using this feedback dataset.

The effectiveness of LLM feedback, and of LLM generated outputs more broadly, can be undermined by undesired and inconsistent behaviors, including hallucination, unfaithful reasoning, and toxic content. A promising approach to rectify these flaws is using self-correction or self-improvement techniques, in which a source of automated feedback, either produced by the LLM itself or some external system, can prompt or guide the LLM to fix problems in its output (Pan et al., 2023). Self-correction methods can be categorized into training-time, generation-time, and post-hoc corrections. Our self-improvement method is most related to training-time self-corrections. For example, Huang et al. (2023) used self-consistency (Wang et al., 2023b) and chain of thought (CoT) prompting to select best generations for further supervised fine-tuning (SFT) on reasoning tasks. Ye et al. (2023) fine-tuned LLama models with self-feedback and revision data generated by ChatGPT to enable the model to self-revise its outputs. Concurrent to our work, Yuan et al. (2024) uses iterative LLM-as-a-Judge (Zheng et al., 2023) prompting to obtain self-rewards and perform direct preference optimization (Rafailov et al., 2023) to perform model alignment to the preferences from this self-reward.

In our work, undesirable and inconsistent LLM feedback generation may include poor goal identification or utterance-level rewrites that are inconsistent with the conversation goals. To mitigate this, we developed a training-time self-improvement method that relies on the fine-tuned LLM itself to provide automated scoring feedback on candidate outputs; this allows it to select preferred generations upon which the feedback model can be further preference-tuned.

3 Feedback Framework

Given the crucial role of human supervision and tailored contextual feedback in the peer counselors training process (Borders and Brown, 2005; Bernard and Goodyear, 1998; Gonsalvez and Milne, 2010; Rønnestad and Skovholt, 2013), we collaborated with senior psychotherapy supervisors (each with over 20 years of experience) to develop an automated feedback system that is aligned with best peer counseling practices. Together, we co-designed a multi-level feedback framework for peer counselor training.

Four one-hour co-design sessions with these senior supervisors revealed that initial training of novice therapists emphasizes foundational active listening skills and that these are generic skills common to all therapy approaches, including peer counseling (Watkins Jr and Milne, 2014; Laska et al., 2014; Wampold, 2015; Cuijpers et al., 2019). Details about the co-design process including research questions, key themes, and the outcomes are given in Appendix B. Via our co-design, we found that the structure of the supervisors' feedback spans different levels: it often starts with positive reinforcement, followed by a line-by-line analysis of session transcripts; for any utterances needing improvement, supervisors clarified the session goals, identified categories of skills that could be improved, and suggested alternative responses.

3.1 Multi-Level Feedback Taxonomy

Building upon our co-design sessions, we derive a multi-level feedback framework that reflects the components of senior psychotherapy supervisors' feedback and trains foundational listening skills that are relevant to peer counseling; see Figure 1. This taxonomy has five key components:

1. **Appropriateness** indicates whether a peer counselor's response in a given context is appropriate and aligned with foundational active listening best practices. No further feedback will be provided if the response is appropriate.

2. **Goal and Alignment**. Unlike casual conversations, peer counseling is goal-oriented, with each question or statement purpose-driven. This component defines what the counselor's goal in this part of the conversation should be and how the response can be changed to improve the alignment to this goal.

3. **Areas for Improvement**. Re-iterating with domain experts and consulting mental health literature (Hill, 2009; 7Cups, 2023), we identify eight widely-used categories of effective communication for peer counseling context: *Reflections, Questions, Suggestions, Validation, Self-disclosure, Empathy, Professionalism, Structure*. Areas of improvement highlights a set of categories that counselors need to further improve.

4. **Alternative Goal-Aligned Response** suggests an alternative response that aligns with the predefined goals and improves over these highlighted areas that need improvement, for a given context.
5. **Positive Reinforcement** (optional) highlights a set of concrete categories as defined in *Areas for Improvement* the peer counselors excel at.

Our multi-level feedback taxonomy, co-designed with senior psychotherapy supervisors, is the first of its kind to resemble how supervisors deliver feedback to counselors post-session. Unlike previous methods that only did one or the other, it uniquely combines evaluating responses and suggesting alternatives. Furthermore, the goal and alignment is a unique component of the taxonomy which explains how to improve alignment to a session-level goal.

FeedbackESConv		
Number of sessions	400	
Number of utterances	8179	
Number of appropriate utterances	4721	(57.7%)
Number of inappropriate utterances	3458	(42.3%)
Avg. length of alternative response	28.3	
Avg. length of goal alignment	36.6	
Categories	-	+
Reflections	616	831
Questions	1431	1995
Suggestions	1159	259
Validation	901	1774
Self-disclosure	558	614
Empathy	1185	3313
Professionalism	279	462
Structure	333	1030

Table 2: FeedbackESConv: Statistics describing the number and average length of feedback annotations at different levels, as well the breakdown of highlighted categories for Areas of Improvement (-) and Positive Reinforcement (+).

4 *FeedbackESConv* Dataset

In order to develop an automatic model that provides contextualized feedback at multiple levels, we use the feedback taxonomy to annotate peer counseling conversations. Given the sensitive nature of peer counseling data and the involved ethical implications, we chose a publicly available counseling dataset *ESConv* (Liu et al., 2021) as our starting point, which contains a large number of emotional support conversations. ESConv was collected on a crowd-sourcing platform, thus requiring quality control. We performed a manual review to filter out conversations that were either low quality or irrelevant to peer counseling (refer to Appendix C for the comprehensive filtering criteria). We divided the obtained dataset into three parts: a dataset with 400 conversations for further annotation by domain experts; a dataset of 150 conversations (Preferences QESconv) used for obtaining self-scored preference pairs as described in Section 5; and a test dataset of 67 conversations.

4.1 Domain Experts

To obtain high-quality annotation, we take a user-centered approach by working with domain experts who have mental health expertise and hand-on practice experience. We recruited domain experts from the Upwork platform by using a selective hiring process (see Appendix D for the hiring criteria). Our final annotator group consisted of two experts who cross-validated the quality of each others' work (see Appendix G.1) – both with over 10 years of experience in professional mental health practice (one was a *Certified Chemical Dependency Counselor* and the other an *Associate Professional Clinical Counselor*).

4.2 Model-in-the-loop Co-annotation

Recent work has shown that LLMs can offer a certain amount of facilitation for data annotation (Li et al., 2023). Thus, to facilitate the annotation process, we leverage a *model-in-the-loop annotation paradigm*, with GPT-4 and domain experts working together on the annotation task – the approach we later refer to as GPT-4+Expert.

Before doing so, we rigorously compare the effectiveness of this co-annotation paradigm, where we set up a comparison of two approaches: generation of initial pre-annotations by GPT-4 and the subsequent refinement by experts, and annotations solely produced by experts. A full GPT-4 based annotation was technically possible, however, it was impossible to ensure feedback correctness and relevance without human supervision.

We compare expert-only annotations to GPT-4+expert annotations, revealing that the average score and consistency of annotation quality improved when GPT-4 was used for pre-annotations. Our results (see Appendix G) show that in 80.8% of cases, feedback created with GPT-4 pre-annotations is either preferred by experts (61.1%) or there is no strong preference either way (19.7%). This demonstrates the domain expert's preference for the model-in-the-loop co-annotation paradigm. We hypothesize that this is because it allows experts to focus on what is most important and refining parts where GPT-4 failed, which we notice in qualitative analysis. As a result, during the annotation process, we use GPT-4 for the initial feedback annotation and then ask our experts to re-work these annotations.

We prompt (see Appendix I) GPT-4 with detailed definitions of each of the feedback components (defined in Section 3.1) and provide in-context examples containing feedback discussed with senior psychotherapy supervisors. We provided domain experts with a detailed annotation guide with definitions and examples of each feedback component as described in our multi-level feedback taxonomy, to get them familiar with the task.

This co-annotation produces annotations of over 400 emotional support conversations. We provide the detailed dataset statistics with the breakdown of highlighted categories for Areas of Improvement (-) and Positive Reinforcement (+) in Table 2.

5 Model

We leverage the resulting FeedbackESConv dataset to develop models that can generate contextualized feedback at different levels for peer counseling. To enable transparent model development, we build upon the open-source Llama-2 model and introduce a simple but effective self-improvement method to generate multi-level feedback.

The open-source approach is crucial in our domain application due to data sensitivity concerns, as it allows the model to be hosted in-house by data owners, ensuring compliance with data storage and analysis regulations. Moreover, an open-source model provides a cost-effective solution that scales well to large mental health communities, circumventing the significant expenses associated with relying on third-party API calls, such as those required by GPT-4.

5.1 Problem Definition

Formally, we define the task of feedback generation based on our multi-level feedback framework as: (1) given the peer counselor's utterance U_i and a context of the peer counselor-seeker conversation, decide if the peer counselor's response is appropriate or needs further improvement by setting y_i to $\mathtt{true}$ or $\mathtt{false}$, respectively. (2) If the response is classified as needing improvement, provide goal and alignment $goal_i$ (text), areas for improvement $ar-_i$ (list), and an alternative goal-aligned response A_i (text). (3) Optionally, provide positive reinforcement or good areas $ar+_i$ (list) for this utterance as a form of positive reinforcement. We represent the feedback generation model as $\mathcal{M}$.

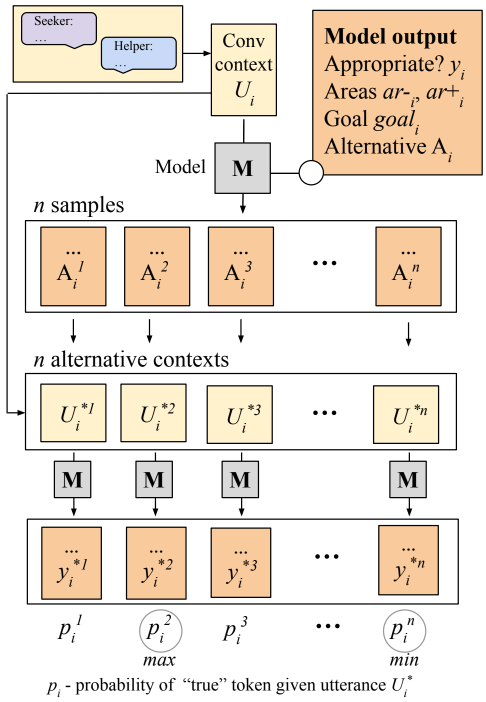


Figure 2: Illustration of the self-scoring mechanism – Phase 1 of the self-improvement method. The first step is to generate n alternative answers for a given conversation utterance U_i. By substituting an alternative answer for the original utterance and passing it back to the model we obtain the probability of the alternative answer being marked as *appropriate*. These scores can be used to create preference pairs for further alignment.

5.2 Self-improvement via Forecasting

The specifics of our multi-level feedback framework allow us to suggest a self-improvement method for $\mathcal{M}$ that does not require any teacher model or additional costly expert data annotation.

On a high level, we take advantage of the fact that both response quality assessment (y) and alternative answer (A) are part of our feedback taxonomy. By substituting an alternative answer for the original utterance, our method uses the feedback model once again to forecast how generated alternative answers will be assessed. This forecast operation estimates the quality of the originally generated feedback and can then be used to guide further alignment of the model. This self-improvement method has the potential to generalize to other scenarios since it applies to any model that jointly assigns binary y_i ($\mathtt{false}$ or $\mathtt{true}$) label and suggests improvements for $y_i = \text{false}$.

Concretely, to enable the self-improvement method with forecasting, we create self-scored preference pairs of feedback generations. To achieve that, we first establish a self-scoring function (Phase 1) and then use sample generations to choose the ones with maximum and minimum scores to form a pair (Phase 2). The model is then aligned to those self-scored preferences (Phase 3).

Phase 1: Self-scoring The goal is to establish a self-scoring function. Our feedback framework is designed in such a way that an alternative answer A_i is part of the output of the model $\mathcal{M}(U_i)$. Hence, we can feed back the alternative answer A_i to the original utterance U_i and substitute it for the originally provided answer and obtain U_i^* (Figure 2). This constitutes a self-assessment loop because we can evaluate the quality of U_i^* by once again passing it to $\mathcal{M}$. The proposed score is the probability of obtaining feedback labeled as *appropriate* ($y_i = true$) for the refined utterance U_i^*. In summary, a feedback generation is assigned a high score if after following the advice (i.e. modifying the peer counselor's response in the suggested way) the probability of $y_i = true$ is high for this altered context. This self-scoring mechanism is a proxy of feedback quality, as we assume that good feedback will lead to good alternative answers.

Phase 2: Preference Pairs Building on the self-scoring mechanism from Phase 1, these self-scores are obtained for a set of samples of $\mathcal{M}$ for the same utterance U_i. Samples with the maximum and minimum scores are indexed with ω_i and α_i, respectively. If the probability that the original utterance U_i receives feedback labeled as *appropriate* is below 0.5 (indicating that further improvement is required), a preference pair is formed using samples ω_i and α_i.

As a robustness check to assess whether these preference pairs are aligned with human judgment, we asked domain experts (see Appendix D for the hiring criteria and Appendix E, G.1 for quality cross-validation studies) to annotate 20 test conversations with minimum and maximum score samples. They preferred the utterance with the higher score 63.0% of the time, had no preference 28.9%, and only preferred the utterance with the lower score 8.1% of the time.

Phase 3: Alignment The last step is to further align the model with Direct Preference Optimization (DPO) (Rafailov et al., 2023) to the preference pairs obtained from Phase 2. This technique contrasts high and low-quality generations and encourages the model to produce generations similar to the ones marked as preferred. We align $\mathcal{M}$ on the Preferences QESconv dataset introduced in Section 4. The resulting model is $\mathcal{M}_{\text{self-imp}}$.

5.3 Baselines

$\mathcal{M}_{SFT}$ *baseline*[2]. To evaluate the self-improvement via the forecasting method, we compare it with a supervised fine-tuned Llama2 13B model baseline, denoted as $\mathcal{M}_{\text{SFT}}$. To understand whether the different phases in the self-improvement method are essential, we compare it with two additional baseline ablation conditions:

$\mathcal{M}_{SFT}$ + *new data*. We apply the $\mathcal{M}_{\text{SFT}}$ model to obtain feedback generations for the additional data Preferences QESConv that $\mathcal{M}_{\text{self-imp}}$ uses. We use those generations for further supervised fine-tuning. The goal here is to determine if self-scoring gives value beyond simply fine-tuning on additional generations on new data used by $\mathcal{M}_{\text{self-imp}}$.

$\mathcal{M}_{SFT}$ + *best scores*. We follow the self-scoring procedure, but instead of creating a single preference pair, we generate multiple scored samples and choose the one with the highest score for further fine-tuning the $\mathcal{M}_{\text{SFT}}$ model. The aim is to see whether alignment to preference pairs gives improvement compared to fine-tuning to the highest-scored generation.

6 Evaluation and Results

In Section 6.1, we compare the quality of feedback generated with $\mathcal{M}_{\text{self-imp}}$ vs. those generated with baseline models via automatic scores and domain-expert ratings. After validating the improved feedback quality of $\mathcal{M}_{\text{self-imp}}$ over baselines, in Section 6.2, we compare its feedback to the feedback co-annotated by GPT-4+Experts (approach described in Section 4.2) to understand if the $\mathcal{M}_{\text{self-imp}}$ model matches in quality.

6.1 Comparing $\mathcal{M}_{\text{self-imp}}$ with Baselines

We use the automatically-computed quality scores (as defined in Section 5.2) as one way to evaluate the performance of our self-improvement method against baselines[3]. For each model, we generate 10

[2] Training details can be found in Appendix H.

[3] To ensure a fair comparison, we perform scoring using the same base $\mathcal{M}_{\text{SFT}}$ model.

Method	$\mathcal{M}_{\text{SFT}}$	$\mathcal{M}_{\text{SFT}}$ + new data	$\mathcal{M}_{\text{SFT}}$ + best scores	$\mathcal{M}_{\text{self-imp}}$
Mean Score Overall	0.968	0.967	0.971	**0.983***
Mean Score Worst 1%	0.28	0.28	0.38*	**0.56***
Mean Score Worst 5%	0.64	0.64	0.69*	**0.81***

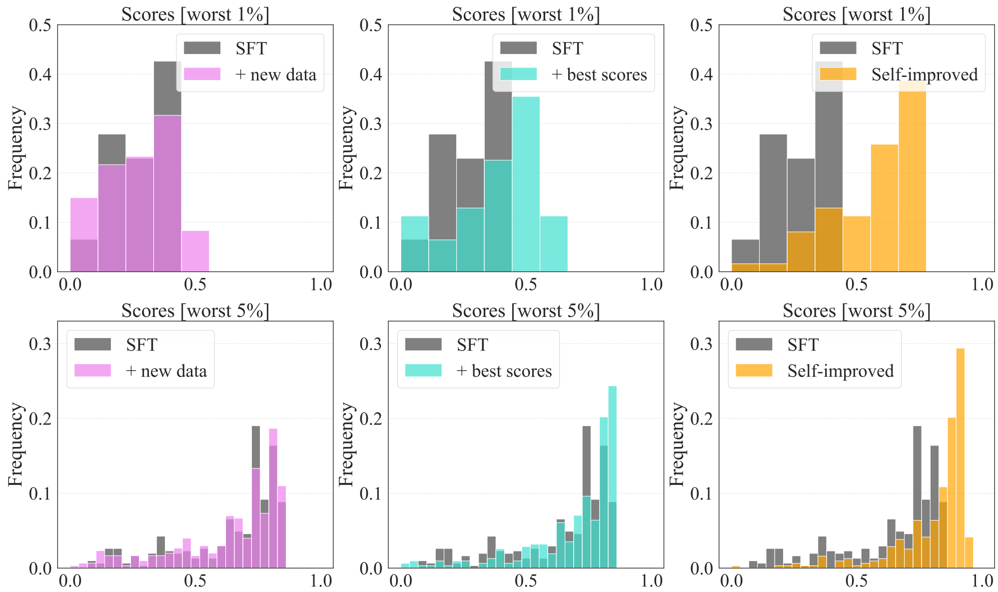


Figure 3: Baselines comparisons. The table presents means of automatically computed quality scores (as defined in Section 5.2) for three baselines and the self-improvement method. The comparison is shown for three different groups: overall and for the worst 1% and 5% of the generations. * denotes statistically significant ($p < 0.01$) improvements over the $\mathcal{M}_{\text{SFT}}$ baseline based on the t-test and Mann–Whitney U test. Plots present score distributions.

samples of feedback for each counselor utterance in 67 test conversations resulting in 8090 data points. Our results are reported in Figure 3. Over all feedback generations, the mean quality score is highest for $\mathcal{M}_{\text{self-imp}}$, where the difference compared to $\mathcal{M}_{\text{SFT}}$ is statistically significant.

In the context of peer counseling, unlike typical natural language generation tasks where average performance is key, our focus is on minimizing the chance of producing poor or unhelpful feedback, prioritizing the worst-case scenario. We illustrate this with an example of both low-quality and high-quality feedback in Figure 4.

As shown in the table in Figure 3, in the worst 5% and 1% of generated feedback, the quality scores for the $\mathcal{M}_{\text{self-imp}}$ model are significantly higher than the baselines. For the bottom 1% of samples, the mean score increases from 0.28 for $\mathcal{M}_{\text{SFT}}$ to 0.56 for $\mathcal{M}_{\text{self-imp}}$, indicating a reasonable shift from inappropriate to appropriate feedback.

Automatically computed quality scores enable observations of improvements on the aggregate distribution level. To affirm that our proposed method $\mathcal{M}_{\text{self-imp}}$ enhances the quality of feedback in the worst-case scenario, we defer to the gold standard of evaluation: the judgment of domain experts.

We conducted the following experiment. We asked domain experts to rate the feedback quality of the bottom 1% of generations using a 5-point Likert scale for $\mathcal{M}_{\text{SFT}}$ and $\mathcal{M}_{\text{self-imp}}$. As shown in the bottom of Figure 5, generations rated as *Very Poor* were almost all eliminated by the use of the $\mathcal{M}_{\text{self-imp}}$ method, to less than 1% of the ratings. Moreover, we see consistent growth of the proportion of generations marked as *Acceptable*, *Good* or *Very Good*. One author further conducted a qualitative investigation of the worst 1% of feedback. We observe that feedback from $\mathcal{M}_{\text{SFT}}$ can often suggest alternative answers with slight rephrasing that do not resolve the core issue, whereas $\mathcal{M}_{\text{self-imp}}$ exhibits fewer of these errors.

Together, the results from these two experiments suggest that for the worst generations, $\mathcal{M}_{\text{self-imp}}$ improves feedback quality as measured both by automatically computed quality scores and domain expert ratings.

6.2 Comparing $\mathcal{M}_{\text{self-imp}}$ with GPT-4+Expert

We further assessed feedback quality at the conversation level and compared feedback generated by

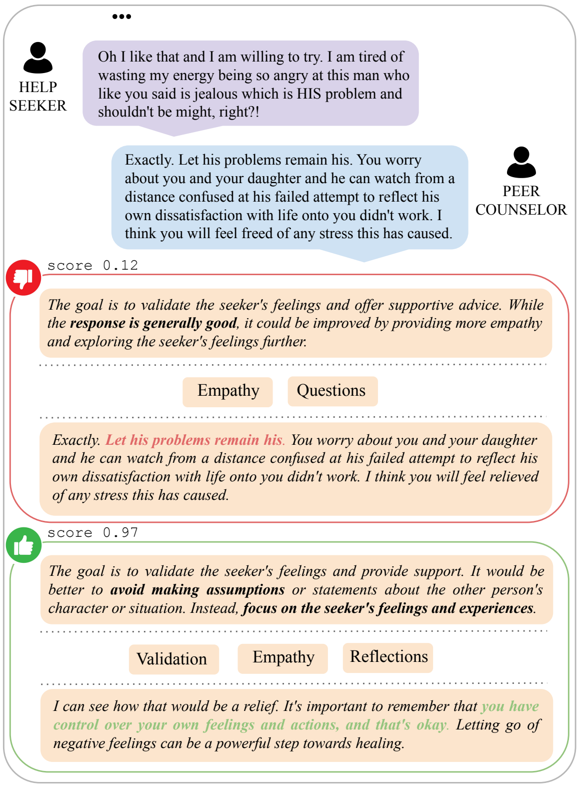


Figure 4: Example of feedback response of very poor quality. The model incorrectly provided feedback that the peer counselor response is *generally good*. Although the model properly outlined the intended goal of the reply, the proposed alternative fails to align with this goal and repeats the same errors. A representation of what constitutes high-quality feedback generation for this specific instance is provided for clarity.

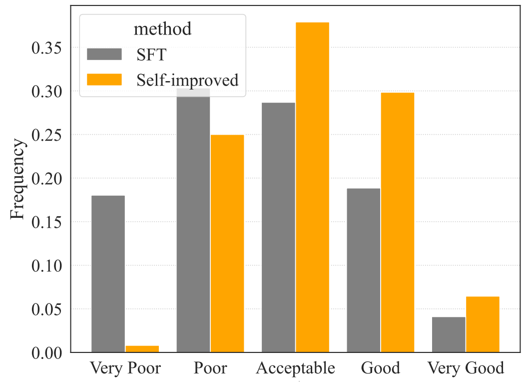


Figure 5: Expert quality assessments for the worst 1% of generations. The statistically significant shift of scores to the right ($p < 0.01$) shows the *self-improvement* method was judged to be of higher quality than the $\mathcal{M}_{\text{SFT}}$ baseline, with mean score improving from 2.61 (Below Acceptable) to 3.16 (Above Acceptable).

Feedback Aspect	$\mathcal{M}_{\text{self-imp}}$	GPT-4 + Expert
Selection for Feedback	4.20	4.18
Strengths Identification	3.68	3.95
Improvement Areas Selection	4.28	4.3
Goal Description Quality	4.3	4.43
Rationale for Alternatives	4.33	4.45
Quality of Alternatives	4.03	4.38*
Feedback Style	4.45	4.55
Feedback Helpfulness	4.15	4.48*
Overall	4.10	4.35*

Table 3: Experts' conversation level evaluation of eight aspects of feedback quality for $\mathcal{M}_{\text{self-imp}}$ and the reference GPT-4+Expert annotations. Results based on a test sample of 20 conversations. * denotes statistically significant difference under t-test ($p < 0.05$).

$\mathcal{M}_{\text{self-imp}}$, against the GPT-4+Expert annotations. Domain experts evaluated the quality of feedback along eight aspects that cover the components of the multi-level feedback taxonomy. Results (Table 3) indicate that the $\mathcal{M}_{\text{self-imp}}$ model's feedback quality approaches the reference standard of GPT-4+Expert annotations across 6 out of 8 feedback aspects, with a median overall quality rating of *4 - Good*. We note significant differences in the *Quality of Alternatives* and overall *Feedback Helpfulness*. Nevertheless, we find that experts agree (4 or 5 on the Likert-scale) in 90% of conversations that the feedback generated by $\mathcal{M}_{\text{self-imp}}$ would be helpful in the training process of novice peer counselors (100 % of GPT-4 + Experts annotations are considered helpful).

These results validate how $\mathcal{M}_{\text{self-imp}}$, a model based on Llama-13B trained using our self-improvement method, can match the GPT-4+expert reference annotations across many aspects while highlighting aspects of the multi-level feedback taxonomy that future modeling work can improve. Example feedback generations are in Appendix J.

7 Conclusions

We introduced a multi-level feedback framework for training counseling skills by co-designing with senior psychotherapy supervisors, constructed a public dataset of counseling conversations with feedback annotations, and proposed a simple but effective self-improvement method for feedback generation. We demonstrate through extensive qualitative and quantitative evaluation that our method minimizes the risk of low-quality feedback generation and generates feedback that domain experts find useful. This work holds the potential to improve the quality and effectiveness of counseling skill training through LLMs.

Limitations

In this work, we first co-designed with senior psychotherapy supervisors a feedback framework and then developed an LLM model that can automatically generate advice for novice peer counselors. Although the framework covers multiple aspects of active listening, it is not enumerative and might not cover all possible feedback dimensions relevant to the complex peer counseling context.

While we consider the way in which the feedback is delivered (and specifically evaluate the feedback style – whether it was delivered "in a friendly but professional way"), we do not tailor our feedback to a specific trainee in a personalized way. In professional training of therapists, supervisors alter their feedback style to optimize feedback delivery: *"But in addition I have a take on who is this person I'm supervising. And what are they like as a person? And do they listen to me or not? And how can I say it differently so they can hear it?"*

Our feedback dataset, which we used for training of our model, was built on a public dataset of emotional support conversations. This allows us to make our data publicly available. However, it was built upon conversations between crowd workers who have only received very abbreviated training. While the training covers a broad range of counseling skills, it is unclear whether these crowd-sourced conversations might generalize to conversations among peer counselors and seekers or other similar counseling contexts.

Although we involved human experts (senior psychotherapy supervisors and domain experts with counseling expertise) at every stage of the development process and system evaluation, we acknowledge that the opinions and judgments from this small group of domain experts might not represent a broader population of psychotherapy supervisors or mental health practitioners, as well as the ways in which they coach novice peer counselors.

We acknowledge that the domain complicates standardized human evaluation due to high subjectivity and resulting disagreements or biases among raters. All the trends presented based on human evaluation also hold on the per-annotator level despite variances in ratings. Future work might consider developing a detailed quality rating system based on the inclusion of a larger group of psychotherapy experts in the discussion.

While automatic evaluation metrics can provide a more objective benchmark for comparing different models, the diverse nature of acceptable feedback in this domain, as confirmed by our co-design sessions with senior supervisors, poses challenges for their application. The development of suitable automatic evaluation methods would require a large annotated dataset with a wide range of expert-generated feedback for each scenario, which was not feasible within the scope of this study due to resource constraints.

Ethics Statement

This study has been approved by the Institutional Review Boards (IRB) at the authors' institutions. All the researchers involved in this study have completed CITI Program certifications on responsible code of conduct in research. We have compensated domain experts fairly for their time, going beyond minimum wage in the United States.

The purpose of this paper is to develop a model that generates feedback for novice peer counselors with limited or no access to human supervision. The system should not be regarded as a substitute for expert feedback. Importantly, while our self-improvement method aims to limit the risk of poor feedback generations (e.g., giving inappropriate advice), this risk is not fully eliminated. It is therefore important to treat model-generated advice only as potential guidance and discard it if necessary, based on trainee judgment.

For potential uses of this feedback generation system, we will design a consent form to disclose potential risks of our system, and will also advocate for practitioners to centrally host and log the content generated by our system so that it can be audited to determine whether there are any problematic behaviors in the system use.

8 Acknowledgments

We would like to thank Camille Harris, Minzhi Li, Michael Ryan, Omar Shaikh, Weiyan Shi, Chenglei Si, Rose Wang, Yanzhe Zhang, Caleb Ziems, and all the members of SALT Lab for their helpful feedback and comments on the manuscript. We also thank the anonymous reviewers for their time and suggestions. This work is supported in part by an NSF grant IIS-2247357 and by the Stanford Impact Labs award.

References

7Cups. 2023. 7cups verfiers team mock chat guide: Discussing points that need improvement.

Kathina Ali, Louise Farrer, Amelia Gulliver, Kathleen M Griffiths, et al. 2015. Online peer-to-peer support for young people with mental health problems: a systematic review. *JMIR mental health*, 2(2):e4418.

Kyle Arnold. 2014. Behind the mirror: Reflective listening and its tain in the work of carl rogers. *The Humanistic Psychologist*, 42(4):354–369.

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

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Janine M Bernard and Rodney K Goodyear. 1998. *Fundamentals of clinical supervision*. Allyn & Bacon.

L DiAnne Borders and Lori L Brown. 2005. The new handbook of counseling supervision.

James FT Bugental, J Fraser Pierson, and Kirk J Schneider. 2001. *The handbook of humanistic psychology: Leading edges in theory, research, and practice*. Sage Publications.

Andrew C Butler, Namrata Godbole, and Elizabeth J Marsh. 2013. Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology*, 105(2):290.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *arXiv preprint arXiv:2401.00820*.

Freddy YY Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 26–33.

David Cooper, Keong Yap, Maureen O'Brien, and India Scott. 2020. Mindfulness and empathy among counseling and psychotherapy professionals: A systematic review and meta-analysis. *Mindfulness*, 11:2243–2257.

Pim Cuijpers, Mirjam Reijnders, and Marcus JH Huibers. 2019. The role of common factors in psychotherapy outcomes. *Annual review of clinical psychology*, 15(1):207–231.

Robert W Day and Richard T Sparacio. 1980. Structuring the counseling process. *Personnel & Guidance Journal*, 59(4).

Anna Fang, Wenjie Yang, Raj Sanjay Shah, Yash Mathur, Diyi Yang, Haiyi Zhu, and Robert Kraut. 2023. What makes digital support effective? how therapeutic skills affect clinical well-being. *arXiv preprint arXiv:2312.10775*.

Nikolaos Flemotomos, Victor R Martinez, Zhuohao Chen, Torrey A Creed, David C Atkins, and Shrikanth Narayanan. 2021. Automated quality assessment of cognitive behavioral therapy sessions through highly contextualized language representations. *PloS one*, 16(10):e0258639.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Craig J Gonsalvez and Derek L Milne. 2010. Clinical supervisor training in australia: A review of current problems and possible solutions. *Australian Psychologist*, 45(4):233–242.

Jennifer R Henretty and Heidi M Levitt. 2010. The role of therapist self-disclosure in psychotherapy: A qualitative review. *Clinical psychology review*, 30(1):63–77.

Clara E Hill. 2009. *Helping skills: Facilitating, exploration, insight, and action*. American Psychological Association.

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

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Zac E Imel, Brian T Pace, Christina S Soma, Michael Tanana, Tad Hirsch, James Gibson, Panayiotis Georgiou, Shrikanth Narayanan, and David C Atkins. 2019. Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy*, 56(2):318.

Ian Andrew James, Rachel Morse, and Alan Howarth. 2010. The science and art of asking questions in cognitive therapy. *Behavioural and Cognitive Psychotherapy*, 38(1):83–93.

Holtzblatt Karen and Jones Sandra. 2017. Contextual inquiry: A participatory technique for system design. In *Participatory design*, pages 177–210. CRC Press.

Taja Kuzman, Nikola Ljubešić, and Igor Mozetič. 2023. Chatgpt: beginning of an end of manual annotation? use case of automatic genre identification. *arXiv preprint arXiv:2303.03953*.

Kevin M Laska, Alan S Gurman, and Bruce E Wampold. 2014. Expanding the lens of evidence-based practice in psychotherapy: a common factors perspective. *Psychotherapy*, 51(4):467.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy Chen, Zhengyuan Liu, and Diyi Yang. 2023. CoAnnotating: Uncertainty-guided work allocation between human and large language models for data annotation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1487–1505, Singapore. Association for Computational Linguistics.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint arXiv:2310.01783*.

Marsha M Linehan. 1997. Validation and psychotherapy.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Do June Min, Veronica Perez-Rosas, Ken Resnicow, and Rada Mihalcea. 2023. VERVE: Template-based ReflectiVE rewriting for MotiVational IntErviewing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10289–10302, Singapore. Association for Computational Linguistics.

Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2022. PAIR: Prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Theresa B Moyers, Lauren N Rowell, Jennifer K Manuel, Denise Ernst, and Jon M Houck. 2016. The motivational interviewing treatment integrity code (miti 4): rationale, preliminary reliability and validity. *Journal of substance abuse treatment*, 65:36–42.

OpenAI. 2023. Gpt-4 technical report.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Erik Rautalinko, Hans-Olof Lisper, and Bo Ekehammar. 2007. Reflective listening in counseling: effects of training time and evaluator social skills. *American journal of psychotherapy*, 61(2):191–209.

Michael Helge Rønnestad and Thomas M Skovholt. 2013. *The developing practitioner: Growth and stagnation of therapists and counselors*. Routledge.

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. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

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

Somayajulu Sripada, Ehud Reiter, and Lezan Hawizy. 2005. Evaluation of an nlg system using post-edit data: Lessons learnt. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment*, 65:43–50.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gareth Terry, Nikki Hayfield, Victoria Clarke, and Virginia Braun. 2017. Thematic analysis. *The SAGE handbook of qualitative research in psychology*, 2:17–37.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Bruce E Wampold. 2015. How important are the common factors in psychotherapy? an update. *World Psychiatry*, 14(3):270–277.

Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023a. Step-by-step remediation of students' mathematical mistakes. *arXiv preprint arXiv:2310.10648*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

C Edward Watkins Jr and Derek L Milne. 2014. *The Wiley international handbook of clinical supervision*. John Wiley & Sons.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3).

Bo Xiao, Zac E Imel, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. 2015. " rate my therapist": automated detection of empathy in drug and alcohol counseling via speech and language processing. *PloS one*, 10(12):e0143055.

Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. 2023. Selfee: Iterative self-revising llm empowered by self-feedback generation. Blog post.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Evaluation areas

Table 4 presents specific examples of mistakes which peer counselors can make. These are grouped into 8 categories with definitions aligned with mental health literature.

B Interviews with senior experts

To understand the nature of feedback in professional training, we conducted multiple interviews with three senior psychotherapists with over 20 years of direct supervision experience of novice therapists. We first understood the common practices of feedback-giving sessions and then engaged with supervisors on a representative task of providing feedback on a transcript of an emotional support conversation to simulate the process of communicating feedback to a psychotherapist student.

The interviews focus on the following questions, insights from which guided the framework design process:

– R1: What are important skills for novice counselors?

– R2: How are these skills learned and what is the role of feedback in the learning process?

– R3: What is the structure of this feedback?

We transcribed all audio recordings of the interviews. Then, using a thematic coding (Terry et al., 2017) approach, we analyzed the interview transcripts to identify key themes and patterns across the data. We then studied how those inform our research questions.

R1: What are important skills for novice counselors?

Beginner psychotherapy skills involve increasing the depth of self-description of the support seeker's problems. Experienced psychotherapists can perceive nuances and undertones in conversations that beginners might miss.

> *"I think an experienced psychotherapist can hear some, can hear some things or pick up on some things that a novice therapist maybe won't that are between the lines."* (Supervisor 1)

The main objective for beginners is not necessarily about adhering to a particular model but mastering basic foundational skills.

> *"I'm thinking that with the beginning novice therapist it's less the model than sort of basic foundational skills that we, I think, we're trying to teach"*

Our experts often referred to *"Helping Skills: Facilitating Exploration, Insight, and Action"* textbook by Carla Hill, who has devised a system categorizing these essential helping skills.

The initial training phase focuses on foundational listening skills, which are also crucial for peer counselors to master.

R2: How are these skills learned and what is the role of feedback in the learning process?

Early-stage students undergo training in foundational counseling skills like listening, empathy, and asking open-ended questions.

> *"Students in the beginning, they, they take certain classes on what I might call basic foundational counseling skills, how to listen, how to be empathetic, how to, you know, ask open-ended questions. There's a list. You know, there's a list of skills"* (Supervisor 1)

After their first year, novice therapists undergo a practicum experience where their sessions are taped and reviewed for feedback on foundational skills.

> *"At the end of their first year, they go for their first clinical experience. We call it practicum experience. And their sessions are taped, and their supervisor goes over those tapes with them and gives them feedback on their, you know, on how they're doing on those basic skills."* (Supervisor 1)

It's beneficial for novices to bring session transcripts, as these provide clear evidence of their actions and their consequences. These tapes and transcripts allow both the supervisor and the novice to study the impact of the therapist's actions on the patient.

> *"But also I like them to bring a transcript. Because then I can go show them. See what you did here led to this, which led to this, and this is what you should do instead."* (Supervisor 2)

Reflections	This skill involves repeating or rephrasing clients' statements to identify and acknowledge their feelings. This technique helps clarify the client's emotions and encourages them to explore these feelings further.
References:	(Bugental et al., 2001; Rautalinko et al., 2007; Arnold, 2014; Hill, 2009; Moyers et al., 2014, 2016; Pérez-Rosas et al., 2019; Beck, 2020; Shah et al., 2022)
Example mistakes:	Not reflecting, drawing conclusions from the helper's experience without listening to what the seeker is saying and checking it out with them; Making assumptions beyond what was said; Copying the seeker's words exactly; Stating feelings too definitely rather than tentatively (e.g., "you obviously feel X" vs. "I wonder if you feel X"); Becoming repetitive, not varying the format of restatements (e.g., I'm hearing you feel sad, I'm hearing you have some thoughts about X, I'm hearing you ...); Labeling feelings inaccurately; Not capturing the most salient feeling; Reflecting on many feelings at the same time; Being judgmental; Focusing on the feelings of others and not the seeker; Reflecting when the seeker is resistant to expressing feelings and reflection might add more pressure.
Questions	Questions in peer counseling can be formulated either as inquiries (e.g., "How do you feel about that?") or as prompts (e.g., "Tell me more about your feelings on that"), provided to aid the client in understanding or examining their emotions.
References:	(Bugental et al., 2001; Hill, 2009; James et al., 2010; Moyers et al., 2014, 2016; Beck, 2020; Shah et al., 2022)
Example mistakes:	Making questions too focused in situations in which they should be more open-ended; Trying to cover everything instead of focusing on one aspect; Asking questions without a clear intention/goal; Not encouraging expression of feelings; Not exploring the details of the situation the seeker is coming with; Not asking the seeker to check the facts ("tell me what data you have that supports that", "do you have any evidence that you'd be X if you did Y?"); Asking questions without empathy; Asking lengthy or multiple questions at once; Turning the attention to other people instead of the seeker (i.e., asking what person X did, instead of asking how the seeker felt about X's behavior); Asking too many closed-questions interviewing instead of exploring.
Suggestions	This technique involves offering specific directives or advice that clients can apply outside the counseling sessions.
References:	(Bugental et al., 2001; Hill, 2009; Moyers et al., 2014, 2016; Beck, 2020; Shah et al., 2022)
Example mistakes:	Giving too much or premature advice, answers, or solutions; Telling people what to do, giving direct advice "you should"; Imposing beliefs or personal values on seekers; Trying to debate with the seeker and convince them of the helper's point of view.
Validation	Validation goes beyond simply acknowledging a client's feelings. It actively affirms their experiences and perspectives as understandable and worthy of respect, even if the counselor may not personally share their viewpoints.
References:	(Linehan, 1997; Bugental et al., 2001; Hill, 2009; Moyers et al., 2014, 2016; Beck, 2020)
Example mistakes:	Not letting the seeker know that their feelings are normal; Validating invalid (e.g., validating opinions or seeker's biases); Helper not being there, paying attention to what the seeker brings to the conversation.
Self-disclosure	Sharing of personal experiences can create a sense of empathy and connection, reducing the client's feeling of isolation. This approach is balanced to avoid overshadowing the client's emotions or introducing irrelevant personal details.
References:	(Henretty and Levitt, 2010; Bugental et al., 2001; Hill, 2009; Moyers et al., 2014, 2016; Beck, 2020; Shah et al., 2022)
Example mistakes:	Not turning the focus back to the seeker immediately; Making self-disclosure too long or too complex; Disclosing too much information; Talking too much and not letting the seeker talk more.
Empathy	This skill involves understanding the client's emotions and sharing in their experience, offering a sense of being truly seen and heard. This deeper connection allows counselors to guide clients toward self-discovery and provide targeted support.
References:	(Bugental et al., 2001; Hill, 2009; Beck, 2020; Cooper et al., 2020; Sharma et al., 2020)
Example mistakes:	[Empathetic Emotional Reactions] Not expressing warmth, compassion, concern, or similar feelings towards the seeker in situations in which it would be appropriate; [Empathetic Interpretations] Not communicating an understanding of the seeker's experiences and feelings in situations in which it would be appropriate; [Empathetic Explorations] Not making an attempt to explore the seeker's experiences and feelings in situations in which it would be appropriate; Expressing empathy but without maintaining a professional attitude; Expressing sympathy instead of empathy.
Professionalism	Professionalism refers to setting clear boundaries and using appropriate language and communication style.
References:	(Bugental et al., 2001; Hill, 2009)
Example mistakes:	Overusing slang; Being overly professional and formal, which results in robotic-style conversations; Using vocabulary that expresses too much closeness.
Structure	This skill assists the counselor and client in guiding the conversation effectively, ensuring productive use of time, and covering essential topics. A basic structure, while flexible to individual needs, provides both parties with a sense of security and direction.
References:	(Day and Sparacio, 1980; Bugental et al., 2001; Hill, 2009; Moyers et al., 2014, 2016; Beck, 2020)
Example mistakes:	[beginning] Not establishing a collaborative agenda and a friendly emotional rapport; [middle] Having too many topics on the table at the same time, not focusing on the main problem ("keep it simple"); [end] Not summarizing what the person is going to take away from the conversation; [end] Lack of clear, actionable items or insights for the seeker after the conversation.

Table 4: Examples of evaluation areas in peer counseling communication grouped into 8 categories: *Reflections, Questions, Suggestions, Validation, Self-disclosure, Empathy, Professionalism, Structure*.

R3: What is the structure of this feedback?

When providing feedback to the novice therapist, the experts emphasized the importance of positive reinforcement by starting with what the counselor did well.

> *"I generally start out with. What they're doing well"* (Supervisor 2)

> *"Well, I'd say this is pretty good overall, so I'd give positive feedback first."* (Supervisor 3)

They would then gently introduce **areas for improvement**. The two crucial skills are making proper reflections and asking good open-ended questions. However, many other areas were mentioned by the experts as they analyzed the provided conversation transcripts.

> *"paraphrasing is a main thing. It's just a couple of, I think that's an important piece. You ask the person a question or they start and then you just kind of repeat what they say [...] asking open questions is another really good one thing that people learn to do"* (Supervisor 3)

Using transcripts like the discussed one can be an effective teaching tool, prompting the therapist to think of **alternative responses**.

> *"I would teach it by using a transcript*

like this. And then I'd say [...] what other kinds of things can you think of that if I said them to you, you'd be more likely to really sink into what it is you're trying to come and talk about?" (Supervisor 3)

Counseling should be **goal-focused**, each question or statement should have a goal.

"[...] what were your goals right? What were your goals in making these questions or suggestions or statements? And I would have have them try and think about it." (Supervisor 3)

When going back to the transcript, the expert analyzed it line by line, stopping at each of the helper's responses and giving feedback on it.

"Counselor says "she gave you a lot of meeting and filled your time fondly". OK, So she's interpreting his statement rather than pulling out more of his statement." (Supervisor 2)

Crucially, the experts point out that the delivery of feedback should be in a manner that ensures the counselor doesn't feel criticized.

"How do they deliver it so that the therapist can hear it? And how does the therapist work with the patient? There are two communications going on there" (Supervisor 2)

Senior supervisors were compensated $150/hour.

C ESConv filtering

We manually analyze the conversations in ESConv (Liu et al., 2021) (CC BY-NC 4.0 license) and filter the ones that meet the following criteria:

- Conversation not on topic
- Conversation referring in big part to MTurk
- Conversation not serious: making jokes, etc.
- Ungrammatical
- Chatting mostly about the current situation COVID, not a specific problem (i.e., exchanging news, vaccination discussions, etc.)
- Mostly meta-conversation ("sorry, are you there, I have not seen your message")

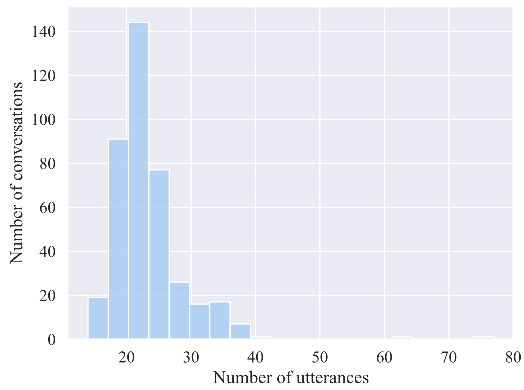


Figure 6: QESConv distribution of the number of utterances in conversations.

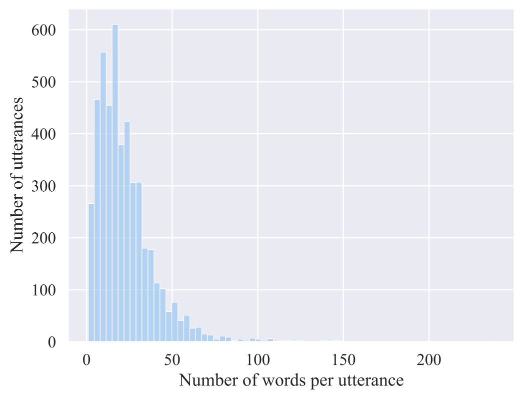


Figure 7: QESConv distribution of the number of words in helper's utterances.

- Generic topic chat: hobbies, having a dog, looking for job advice

In this way we select 400 conversations for the QESconv dataset. We further remove many conversation-finishing artifacts by searching for keywords "survey", "quit", "we need to chat","button" and manually removing those from utterances. For example: *"can you press quit first, I can't do it from my end"* , *"I think we need to chat a bit more in order to wrap things up"*, *"please remember to take the survey :)"*, *"Is there a quit/finish button on your end?"*.

The final dataset has in total 11.3K utterances (distribution shown in Figure 6, with average utterance length equal 21.4 words (distribution for helper in Figure 7 and seeker in Figure 8).

D Domain experts hiring process

Based on the submitted applications and conducted interviews, we choose a group of six experts. We

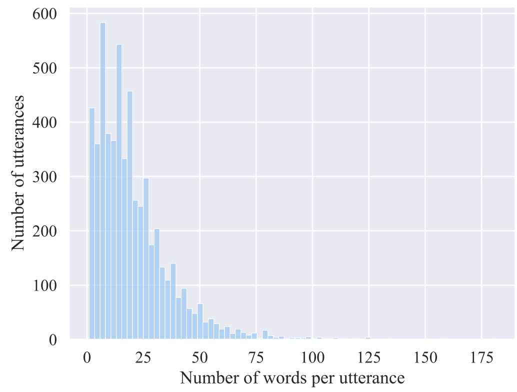


Figure 8: QESConv distribution of the number of words in seeker's utterances.

then conduct a pilot study in which we ask the experts to annotate a single conversation based on our annotation guide describing the feedback framework (Section 3) and our annotating interface. Based on adherence to the guide and projected time availability, we establish a group of three self-validated (at least 4/5 in the Likert scale – for details see Appendix E) experts – all with over 10 years of professional mental health practical experience (for example as *Certified Chemical Dependency Counselor*, *Licensed Marriage and Family Therapist* and *Associate Professional Clinical Counselor*).

Upon further quality tests for the final data annotation scheme, we narrow down the group to two experts who consistently validate the quality of each other's annotations on the final annotation task (see Appendix G.1). Our annotators are US-based.

Domain experts were compensated $30/hour. We informed them of the purpose of the study and the potential risks.

E Pilot quality validation

We observe variability in feedback among experts, but we confirm with senior supervisors that this is to be expected since each practitioner may focus on different counseling components. Since there is no gold truth feedback, [4] evaluating the annotation quality is challenging and requires human expertise. We, therefore, perform a pilot self-validation study in which each expert judged on a 5-point Likert scale the quality of of the other experts' annotations.

In an experiment involving three experts (third expert later excluded at the co-annotation stage), each was tasked with evaluating the annotations made by the others for a single conversation. The assessment was based on a five-point Likert scale:

1. **Completely Irrelevant**: The feedback is unrelated to the task.
2. **Slightly Relevant**: The feedback has minimal relevance, lacking depth or specificity.
3. **Moderately Relevant**: The feedback is partially relevant, covering some, but not all, key aspects.
4. **Highly Relevant**: The feedback addresses most key aspects effectively.
5. **Exceptionally Relevant**: The feedback is comprehensive, insightful, and offers actionable suggestions.

Even though the annotations varied, the experts found different ways of giving valid feedback: *"I think the other annotators and I emphasized things in slightly different ways. For example, one was more focused on clarity and the other was more focused on validation."* They all rated each others annotations to be at least 4/5 validating the overall annotation quality (see Table 5).

All evaluations were blind, i.e. we did not reveal the source of the annotations.

Evaluator	Annotator		
	A	B	C
Expert A	-	4/5	4/5
Expert B	5/5	-	4/5
Expert C	4/5	5/5	-

Table 5: Quality validation pilot results.

F Potential of LLMs for providing feedback

We explore whether LLMs could help in the annotation process within the feedback framework we have defined. This presents a topic of empirical investigation on its own.

LLMs have been used for annotations (Gilardi et al., 2023; Kuzman et al., 2023), or co-annotation (Li et al., 2023), and GPT-3.5 and GPT-4 models excel at classification tasks related to client/therapist behaviors (Chiu et al., 2024); however, our task is

[4] Even identifying areas for improvement cannot be simply defined as a multi-classification problem since different areas can be highlighted and there is not a single correct set

much more open-ended, requiring a generation of natural language rationale using deep understanding of the specialized feedback framework.

We experiment with Llama2-70b chat (Touvron et al., 2023), GPT-3.5 Turbo and GPT-4 models (OpenAI, 2023). While all models give reasonable feedback when prompted with a short generic statement, [5] the feedback is not focused (the most generic for the Llama model).

When provided with a detailed definition of our framework, we find Llama to be unsuccessful in parsing framework guidelines, which both GPT-3.5 and GPT-4 manage to do. However, we find in early experiments that GPT-3.5 produces feedback of significantly inferior quality to human one, therefore we proceed with GPT-4 as our base model, which showed high potential.

F.1 GPT-4 prompting

While the most straightforward approach would be to use an API call to annotate each U_i, it would be very expensive given the usage of the GPT-4 model and the number of tokens in the instruction (>2k). Annotating the full conversation at once would be the most efficient option, but we notice a significant degradation of quality in annotations of the final helper's utterances. Therefore, we annotate overlapping chunks for the conversation of 5 helper's utterances [6].

F.2 GPT-4 quality pilot

Similar to the setting described in in Appendix E, we follow up with GPT-4 quality pilot by annotating ten conversations with GPT-4 and asking the experts for the 5-point Likert scale evaluation (one overall score for ten conversations). The results are presented in Table 6.

Some experts pointed out that sometimes the language seems "stuffy" and "medical", thus leading us to prompt refinement [7] The final prompt can be found in Appendix I.

[5]Example simple prompt: *Act as a supervisor of novice helpers in the mental health context. Give feedback to the helper on their last response in the conversation below.*

[6]The chunks are overlapping; we discard feedback for the first two utterances, which lack the sufficient context.

[7]We refined the prompt additional language consideration: *Use professional and friendly language when giving feedback. Focus on what is most beneficial to hear.*.

Evaluator	Annotator
	GPT-4
Expert A	5/5
Expert B	5/5
Expert C	5/5

Table 6: Quality validation pilot results for GPT-4 generated annotations.

G Expert-only vs. GPT-4+expert annotations

All experts annotated a set of ten conversations. The sets were different so that later evaluations are not biased by comparison to oneself, i.e., "*this is not good because I did something else*"[8]. Additionally, the experts annotated another set of ten conversations, this time, refining GPT-4 feedback.

Each expert then evaluated the quality of annotations made by other experts with and without GPT-4 default feedback (7 questions asking about feedback components, 5-point Likert scale) and compared on utterance level whether expert-only annotation or GPT-4+expert annotation is preferred (or there is no significant difference).

G.1 Do experts consistently validate themselves?

Experts A and B get high ratings, even without GPT-4 pre-annotation. While expert C initially demonstrated the ability to produce high-quality annotations, there appears to be some inconsistency in maintaining the same level of quality across an entire batch of conversations (scores below *Acceptable* rating). Figure 9 presents the average and median score of each expert rated by every other expert. Figure 10 presents how each expert overall (averaged over the raters) was scored in each of the questions asking about different feedback components.

Experts A and B consistently achieve scores around 4 which translates to *Good* quality. Experts C fails to exceed the *Acceptable* rating for all questions across the board. Moreover, their answers are also subject to the highest variation in the score in majority of cases.

[8]Due to the subjective nature of this task, there is no single correct way of annotating.

Average score - Rater to Annotator

Rater	Expert A	Expert B	Expert C
			Annotator / Score
Expert A	-	3.5	2.2
Expert B	3.5	-	2.6
Expert C	4.1	4.1	-

Median score - Rater to Annotator

Rater	Expert A	Expert B	Expert C
			Annotator / Score
Expert A	-	4	2
Expert B	4	-	2
Expert C	4	4	-

Figure 9: Tables presenting average and median score for quality of annotations for experts A, B and C. Each entry in the table shows how a particular expert (row) rated another expert's annotation (column). The summary of each column provides the overall quality score of the expert's annotations.

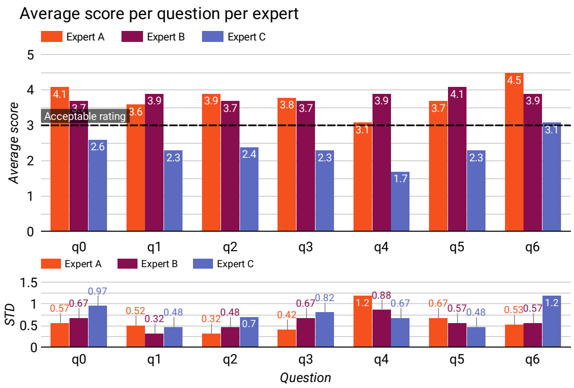


Figure 10: Figure presents average score with standard deviation for each expert A, B, C broken down by 7 questions used to assess the quality of experts annotations. The dotted line marks the *Acceptable* rating (3).

G.2 Do annotations benefit from GPT-4 usage?

With Expert C excluded as a rater, we compare the average annotations quality score of Expert A and B with and without GPT-4 pre-annotations (same setting as in the validation pilot - 7 questions and 5-point Likert scale).

The average score assessing the annotations' quality improves when GPT-4 is used for pre-annotations (Table 7). Moreover, the standard deviation of the scores decreases. Taking these factors combined, the results point to higher and more consistent quality of annotations when GPT-4 is used.

Additionally, GPT-4 + Expert is strongly preferred on the utterance level (see Figure 11). Presented with two annotations, one with and the other without GPT-4 pre-annotations, raters in the majority of cases (61.1 %) prefer the ones with pre-annotations. In 19.7% of cases, they are indifferent, and in 19.3% of cases, they prefer annotations without GPT-4 pre-annotations.

Annotation method	Average score
Expert-only	3.54 ± 0.81
GPT-4 + Expert	3.96 ± 0.62

Table 7: Comparison of the average score of annotations' quality averaged over the experts without and with GPT-4 pre-annotations.

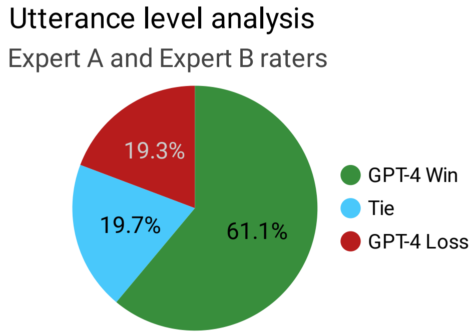


Figure 11: Diagram presenting the distribution of whether the raters (Experts A and B) prefer annotations with or without GPT-4 pre-annotations. The figure presents percentages for three options: GPT-4 Win (green) – 61.1%, Tie (blue) – 19.7%, and GPT-4 Loss (red) – 19.3%.

Qualitatively, when experts refine annotations, they tend to add extra feedback components, for instance, adding an extra goal over the one already pointed out by GPT-4. They sometimes rephrase goal/alternative response chunks that can be improved (e.g., making the question more open-ended). Those were thus not only due to fixing errors but also aim to refine and follow individual preferences (Sripada et al., 2005).

We hypothesize that GPT-4 + Expert are preferred since they allow experts to focus on what is most important and refining parts where GPT-4 failed. This reduces the work burden of writing everything from scratch. Quantitatively, we conduct the Wilcoxon test (Wilcoxon, 1992) on conversation ratings, and statistically, GPT-4+Expert conversations obtain better ratings ($p<0.05$). The win/loss rate is also statistically significant (Wilcoxon and Binomial test, $p <0.05$).

Based on the above pilot results, we continue annotating QESConv with Experts A and B.

H Fine-tuning experimental setup

To curate a fine-tuning dataset, we leverage our FeedbackESConv data. To format each training datapoint we follow Alpaca style instruction formatting (Taori et al., 2023). Each datapoint contains as output the feedback annotations from FeedbackESConv for the utterance U_i, with goal & alignment parts preceding the alternative answer, to provide "explanations" in order to guide the generation process (Wei et al., 2022).

The input is the conversation context c_i. To find the part of the conversation to provide relevant context, we follow (Chen and Yang, 2020) by segmenting the conversation using C99 algorithm(Choi, 2000) on utterance embeddings. We embed the utterances using HuggingFace (Wolf et al., 2020) transformer model *all-MiniLM-L6-v2*. We define the relevant context for each utterance as all past utterances in the current and previous segments.

For the supervised fine-tuning stage, we use the standard causal language modeling objective (cross-entropy on token logits). We fine-tune for three epochs. In the DPO stage, we use the objective from (Rafailov et al., 2023) with the beta parameter set to 0.5. We use a single A100 GPU for the experiments. Overall, our computational budget amounted to approximately 130 GPU hours.

I GPT4 Prompt & In-context Learning

```
In emotional support conversations, two primary roles exist: the helper
(individual providing support) and the seeker (individual seeking support).
Your task is to provide feedback to the helper on these conversations.

## Instructions

Annotate helper’s responses. Give your feedback on all the helper’s responses.

There are two options for annotating the helper’s responses:

Option I
If you believe the response is very good, you can annotate it by setting
perfect key to true.
Please highlight good areas in which the helper excelled.

Option II
Most of the helper’s responses could be improved with your feedback, you can
annotate it by setting perfect key to false.
If there are any particularly good areas in the responses (even though the
response has mistakes related to other areas), please highlight them.
The feedback should have three different parts. Give constructive feedback to
the helper consisting of the following three parts A, B and C:

### Part A (feedback)
What should the goal of this response be? What could I [helper] improve to
better align my response with this goal? Always start with "The goal is to" and
then specify the goal for this part of the conversation.
Think about the context, what is the most important goal in this part of the
conversation? Potential goals could be defining the seeker's emotions or a
problem, identifying possible causes of the seeker's problems, helping the
seeker identify helpful changes, helping the seeker understand what thoughts
they have that do not help them, etc. Please formulate the goals yourself, if
there is more than one goal, pick the most important one.

[Important] Structure your response in the following way:
Start with “The goal is to,” then specify the goal for this part of the
conversation.
Then, say what could be improved to achieve this goal. Please use third-person
statements “it would be good to”, “it might be better to”, “it would be great
to”, etc. (This is to ensure good feedback delivery)

### Part B (areas)
What areas for improvement do you want to highlight?
Identify the categories of improvement from the list provided (see
Appendix:Areas for improvement). The list contains the areas in which novice
helpers struggle. Please study it in detail to understand each category.

### Part C (alternative)
Give a potential alternative response I [helper] could have given.
```

Offer more suitable responses, the helper could have used in the context of the conversation in order to achieve the goal specified in part A.

While annotating, think about the whole context of the helper’s response. What happened earlier in this conversation, and how does the helper’s response fit into this context?

Important Language Considerations
Do not say “the helper did.., etc. The feedback should always be delivered using phrases like “it might be”, “it would be better”, “... would be more effective”, etc. Please try not to repeat the same phrase in one annotation.
Always refer to the seeker using gender-neutral terms like "they" unless their gender is explicitly stated.
Ensure your feedback is respectful, objective, and constructive. We do not want to judge the helper but to help them master their skills further.
While giving feedback, please feel free to quote parts of the dialogue using “”, if you find it helpful to refer to what the seeker or helper said.

Appendix: Areas for improvement

Reflections
- Not reflecting, drawing conclusions from the helper’s experience without listening to what the seeker is saying and checking it out with them
- Making assumptions beyond what was said
- Copying the seeker's words’ exactly
- Stating feelings too definitely rather than tentatively (e.g. you obviously feel X vs. I wonder if you feel X)
- Becoming repetitive, not varying the format of restatements (e.g. I’m hearing you feel sad, I’m hearing you feel anxious, I’m hearing you…)
- Labeling feelings inaccurately
- Not capturing the most salient feeling
- Reflecting on many feelings at the same time
- Being judgmental
- Focusing on the feelings of others and not the seeker
- Reflecting when the seeker is resistant to expressing feelings and reflection might add more pressure

Questions
- Making questions too focused in situations in which they should be more open-ended
- Trying to cover everything instead of focusing on one aspect
- Asking questions without a clear intention/goal
- Not encouraging expression of feelings
- Not exploring the details of the situation the seeker is coming with
- Not asking the seeker to check the facts (tell me what data you have that supports that”, do you have any evidence that you'd be X if you did Y?)
- Asking questions without empathy
- Asking lengthy or multiple questions at once

- Turning the attention to other people instead of the seeker (i.e., asking what person X did, instead of asking how the seeker felt about X’s behavior)
- Asking too many closed-questions interviewing instead of exploring

Suggestions
- Giving too much or premature advice, answers, or solutions
- Telling people what to do, giving direct advice “you should”
- Imposing beliefs or personal values on seekers
- Trying to debate with the seeker and convince them of the helper’s point of view

Validation
- Not letting the seeker know that their feelings are normal
- Validating invalid (e.g., validating opinions or seeker’s biases)
- Helper not being there, paying attention to what the seeker brings to the conversation

Self-disclosure
- Not turning the focus back to the seeker immediately
- Making self-disclosure too long or too complex
- Disclosing too much information
- Talking too much and not letting the seeker talk more

Empathy
- [Empathetic Emotional Reactions] Not expressing warmth, compassion, concern, or similar feelings towards the seeker in situations in which it would be appropriate
- [Empathetic Interpretations] Not communicating an understanding of the seeker's experiences and feelings in situations in which it would be appropriate
- [Empathetic Explorations] Not making an attempt to explore the seeker's experiences and feelings in situations in which it would be appropriate
- Expressing empathy but without maintaining a professional attitude
- Expressing sympathy instead of empathy

Professionalism
- Overusing slang
- Being overly professional and formal, which results in robotic-style conversations
- Using vocabulary that expresses too much closeness

Structure
- [beginning] Not establishing a collaborative agenda and a friendly emotional rapport
- [middle] Having too many topics on the table at the same time, not focusing on the main problem (“keep it simple”)
- [end] Not summarizing what the person is going to take away from the conversation

- [end] Lack of clear, actionable items or insights for the seeker after the conversation

Give feedback to all helper's responses. Use professional and friendly language when giving feedback. Focus on what is most beneficial to hear.

Conversation

Helper: Hello. How are you doing today?
Seeker: Feeling pretty down to be honest.
Helper: Oh, I am sorry about that. Why are you feeling down?
Seeker: I'm just really lonely. My friends are all very busy lately and I haven't been able to find a partner for a long time.
Helper: I can understand that. It is difficult feeling alone.
Seeker: Yes. Normally it's not so bad but it's been such going on for such a long time. It's harder to deal with after so many years.
Helper: It sounds like you feel your friends are too busy for you.
Seeker: Yes, but the biggest part is not being able to find romantic partner.
Helper: Why do you think you are having trouble finding a suitable romantic partner?
Seeker: Part because of my low income and part because of my age. I live in a college town and most single women are 10 years younger than me.
Helper: Sometimes meeting people through mutual friends is helpful. Have you asked any of your friends if they could introduce to you people they know?

Annotation


```
{
  "annotations": [
    {
      "helper": "Helper: Hello. How are you doing today?",
      "perfect": true,
      "goodareas": [
        "Structure"
      ]
    },
    {
      "helper": "Helper: Oh, I am sorry about that. Why are you feeling down?",
      "perfect": true,
      "goodareas": [
        "Questions"
      ]
    },
    {
      "helper": "Helper: I can understand that. It is difficult feeling
alone.",
      "perfect": false,
      "goodareas": [
        "Empathy"
```

```
      ],
      "feedback": "The goal is to find which problem is to be addressed. There
are two possible problems: the seeker's friends are busy, and they want to find
a partner. It would be great to ask the seeker which of the two problems is
more important to work on.",
      "badareas": [
        "Questions",
        "Structure"
      ],
      "alternative": "Both of these things sound tough, leading you to feel
alone. Is there one of the two problems that feels more important right now?"
    },
    {
      "helper": "Helper: It sounds like you feel your friends are too busy for
you.",
      "perfect": false,
      "goodareas": [],
      "feedback": "The goal is to validate the feelings and then check the
facts. The seeker shows all-or-nothing, so it would be a good idea to express
empathy but then check if the facts line up with the seeker's comment.",
      "badareas": [
        "Reflections",
        "Validation",
        "Questions"
      ],
      "alternative": "I can tell that the problem has gone for a long time and
it feels overwhelming right now. You said your friend right now are busy What
do you mean by it? Tell more about how long you haven't been seeing friends."
    },
    {
      "helper": "Helper: Why do you think you are having trouble finding a
suitable romantic partner?",
      "perfect": false,
      "goodareas": [],
      "feedback": "The goal is to find out more about the problem with finding
a partner. It would be better to be more open-ended about the causes of that.",
      "badareas": [
        "Questions"
      ],
      "alternative": "Oh, tell me more about what is going on with that."
    },
    {
      "helper": "Helper: Sometimes meeting people through mutual friends is
helpful. Have you asked any of your friends if they could introduce to you
people they know?",
      "perfect": false,
      "goodareas": [],
      "feedback": "The goal is to still assess the problem. Rather than giving
suggestions really early, the response could paraphrase the statement to
```

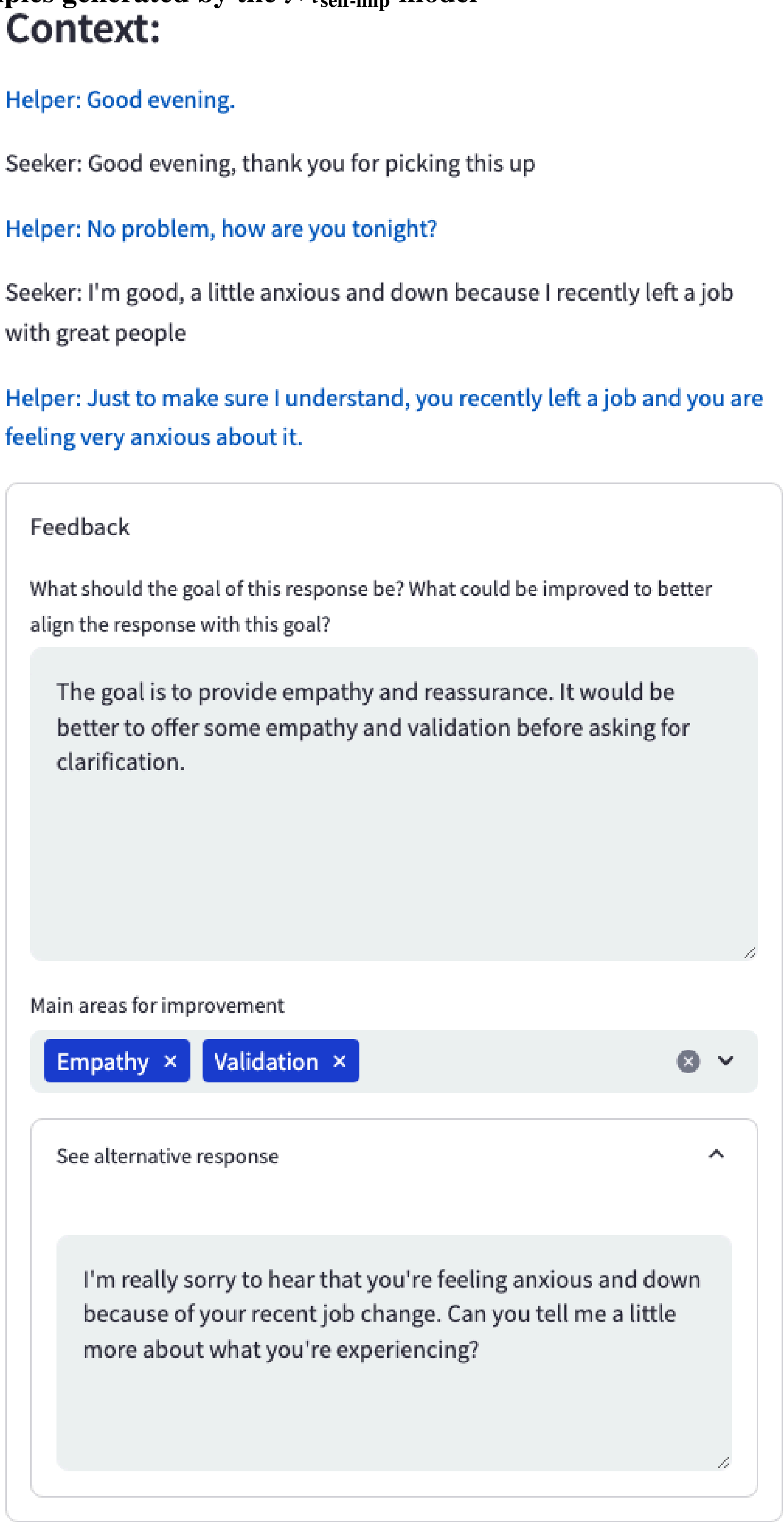
```
encourage the person to share more details and explore the problem a little
more.",
      "badareas": [
        "Questions",
        "Empathy",
        "Suggestions"
      ],
      "alternative": "It sounds like that could get in the way; what else might
be going on?"
    }
  ]
}
```

Conversation

<SEGMENT TO BE ANNOTATED>

Annotation

J Feedback samples generated by the $\mathcal{M}_{\text{self-imp}}$ model



Context:

Seeker: Hi, my work situation is stressful

Helper: Hi, I'm sorry to hear that. What's been happening?

Seeker: They barely acknowledge my presence and don't really talk to me except for an occasional hello

Helper: That's not very fair. You deserve to feel valued at your workplace and it sounds like you don't feel that way.

Seeker: I know, I don't know why they act this way. is it something I have done. I'm stressed.

Helper: You haven't done anything wrong. I have felt that way at work before too. I went to therapy to deal with my difficult work situation.

Seeker: therapy, is it maybe me, or should I bring it up with my boss.

Helper: You could bring it up with your boss. That would be a very brave thing to do.

Feedback

What should the goal of this response be? What could be improved to better align the response with this goal?

The goal is to explore the seeker's thoughts and feelings about their situation before suggesting actions. It would be better to ask more questions about the seeker's feelings and thoughts about their situation, their boss, and their workplace culture before suggesting actions.

Main areas for improvement

Suggestions × Questions ×

See alternative response

It's great that you're thinking about discussing this with your boss. Can you tell me more about how you're feeling about your work situation and your boss? How do you think your boss might react if you bring this up with them?

Context:

Seeker: Hello. I'm looking for some help.

Helper: Good Morning! would you like to tell me whats going on?

Seeker: I have a big test coming up for a college class and I am very anxious about it.

Helper: you seem really concerned about a big test coming up and its making you feel anxious. is that correct?

Seeker: Yeah exactly. I feel like my future depends on getting a good grade on this test.

Helper: just from chatting with you, I can feel how anxious you are.

Seeker: I've been studying every day for weeks but I still can't help but feel really anxious. I don't know what to do. It's messing with my sleep schedule.

Helper: your doing your best and continuing to try hard!. it's great that you are working so hard to pass this test.

Feedback

What should the goal of this response be? What could be improved to better align the response with this goal?

The goal is to acknowledge the seeker's struggles and express empathy. It might be better to avoid making assumptions about the seeker's efforts and instead, ask more open-ended questions to better understand their situation.

Main areas for improvement

Questions × Validation ×

See alternative response

It sounds like you're putting a lot of effort into your studies, and it's still causing you a lot of stress. Can you tell me more about how you're feeling and what's been most challenging for you?

Context:

Seeker: Hello

Helper: Hello!

Seeker: I am not feeling very good about myself lately

Helper: Why are you not feeling very good about yourself, lately?

Feedback

What should the goal of this response be? What could be improved to better align the response with this goal?

The goal is to create a safe and empathetic environment for the seeker to express their feelings. It would be better to avoid asking a question that might sound accusatory and instead show empathy towards the seeker's feeling.

Main areas for improvement

Empathy ×

See alternative response

I'm sorry to hear that you're feeling this way. Can you tell me more about what's been going on?

Context:

Seeker: Hi

Helper: Hello! How can I encourage you tonight?

Seeker: I'm really anxious and I javen. haven't been sleeping lately

Helper: So you have a lot on your mind and you are losing sleep over it?

Seeker: Yeah my company laid off 10% of the workforce last Friday and furloughed another 10%. I'm afraid I'm next

Helper: I'm sorry to hear that. It can be a bit stressful to think you may lose your job.

Seeker: yeah my mind is racing and I can't stay calm.

Helper: I have found there are ways to calm myself down when things in my life seem very uncertain.

Seeker: Really? What seems to work? I'm willing to try anything right now

Helper: I make a list of the things I can do - And the things I can't. Then you can have goals. And if you can work on a goal, for example, if you need to write a resume, you start working on that a little bit every day until it's done. But if there is nothing you CAN do about the situation, you have to pray to God to help you then LET IT GO. If you can NOT fix it now, you will only make yourself sick thinking about it.

Seeker: I know. I can't seem to shut my mind down. And not getting sleep doesn't help

Helper: It is much more difficult for me to remain focused and positive when I lose sleep. I have had to focus on self care, finding a friend to talk about my problems with, not someone who will TELL me what to do, but support me. So you need a good support system. And I find things like prayer, Bible reading, nice soft sheets to sleep on and some herbal tea like Sleepy Time help me relax before bedtime.

Feedback

What should the goal of this response be? What could be improved to better align the response with this goal?

The goal is to provide support and suggestions based on the seeker's specific situation. It would be more effective to focus on the seeker's specific situation and offer suggestions based on their unique circumstances. Also, the self-disclosure might be too long and not relevant enough to the seeker's situation.

Main areas for improvement

Self-disclosure × Suggestions ×

See alternative response

It's understandable that you're feeling this way. Not getting sleep can definitely make things worse. What helps you relax before bedtime?

Context:

Seeker: Yes, but they are rowdy.

Helper: I can see where that would be frustrating

Seeker: I think I will ask my sister if I can use the laundry room as a little office so that I can close the door.

Helper: Have you tried asking their parents to speak with them ? That is a great idea!

Feedback

What should the goal of this response be? What could be improved to better align the response with this goal?

The goal is to validate the seeker's idea before suggesting solutions. It would be better to acknowledge their idea and ask them how they feel about it, rather than jumping to suggesting solutions too quickly.

Main areas for improvement

Suggestions × Validation ×

See alternative response

That sounds like a great idea. How do you think that might help you?