PAIR: Prompt-Aware margIn Ranking for Counselor Reflection Scoring in Motivational Interviewing

Do June Min[1], Verónica Pérez-Rosas[1], Kenneth Resnicow[2], and Rada Mihalcea[1]
Department of Electrical Engineering and Computer Science[1], School of Public Health[2]
University of Michigan, Ann Arbor, MI, USA
{dojmin, vrncapr, kresnic, mihalcea}@umich.edu

Abstract

Reflections are a core verbal skill used by mental health counselors to express understanding and acknowledgement of the client's experience and concerns. In this paper, we propose a system for the automatic evaluation of counselor reflections. Specifically, our system takes as input one dialog turn containing a client prompt likely leading to a reflection and a counselor response to it, and outputs a numeric score indicating the quality of the reflection made by the counselor. We compile a dataset consisting of reflections portraying different levels of reflective listening skills, and propose Prompt-Aware margIn Ranking (PAIR), a novel framework for reflection scoring that contrasts positive and negative prompt and response pairs using adhoc multi-gap and prompt-aware margin ranking losses. Through empirical evaluations and deployment of our system in a real-life educational environment, we show that our scoring model outperforms several baselines on different metrics, and can be used to provide useful feedback to counseling trainees.

1 Introduction

Counselor training is an expensive and time consuming process due to the extensive expert supervision involved (Bartholomew et al., 2007). Current strategies for counselor training usually rely on either role playing or monitoring and recording live video interactions, which are then manually evaluated to provide constructive feedback, thus limiting the opportunities for training counselors to practice and receive timely evaluative feedback.

While several promising approaches have been proposed to automatically provide evaluative feedback to counselors (Tanana et al., 2019; Shen et al., 2020), providing detailed feedback in real time remains a challenge. This is particularly the case in educational settings, where counseling trainees could benefit from a supportive learning environment that allows them to make mistakes and learn at their own pace while acquiring counseling skills.

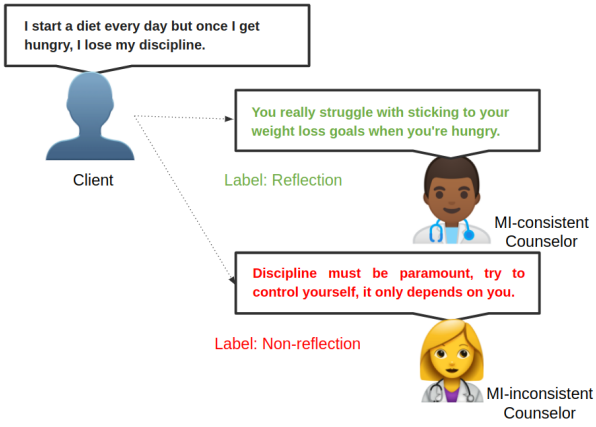

Figure 1: Example of Reflective and Non-reflective Counselor Behaviors.

Seeking to address this need, we introduce the task of scoring the language of counseling trainees when learning to formulate responses to clients' statements. We focus on responses containing counseling reflections, e.g., the ability to understand and reflect on what the client is saying. Our goal is to create a computational model for scoring counselor reflections by leveraging existing counseling behavioral annotation schemes and learning-to-rank approaches. We believe that developing an Natural Language Processing (NLP) model capable of assessing the quality of verbal behavior observed in counselor language can improve the quality and efficiency of Motivational Interviewing (MI) training by allowing counselors to practice in real-time their reflective listening skills, and providing them with immediate feedback.

To build our scoring system, we compile a new dataset of counseling reflections using expert and non-expert annotations. The dataset consists of pairs of client prompts, i.e., situations from a client that prompt a reflective statement, and counselor responses showing different levels of counseling skill, as shown in Figure 1. We also introduce PAIR (Prompt-Aware margIn Ranking), a novel margin

ranking-based approach that can output a continuous score learned from discrete annotations of counseling reflections. We conduct a set of experiments showing that our model is able to learn the correct ranking of counseling responses. In addition to testing the model on the collected data, we deploy our models in a real world education setting with graduate counseling students, and conduct quantitative and qualitative evaluations showing that our system is a viable alternative to manual human feedback.

Our main contributions include: (1) The formulation of the reflection scoring problem and a counseling dataset for this task; (2) A novel contrastive learning-inspired framework PAIR (Prompt-Aware margIn Ranking) for automatically scoring the quality of a reflection statement; (3) Quantitative and qualitative assessments of our model on the annotated dataset and through in-the-wild deployment and feedback.

2 Related Work

Automated analysis and evaluation of verbal strategies used in mental health conversations has emerged as a promising intersection of psychotherapy and NLP (Althoff et al., 2016).

Work has been done to measure the fidelity of treatment via behavioral coding or analysis of counselors' language (Pérez-Rosas et al., 2017b; Flemotomos et al., 2021; Ardulov et al., 2022), and also to evaluate empathy and verbal mimicry expressed by counselors (Pérez-Rosas et al., 2017a; Sharma et al., 2020). More recently, dialog-based systems have been explored to assist the development of basic counseling skills. Tanana et al. (2019) developed a patient-like conversational agent that interacts with counselors while practicing open questions and reflections and categorizes their responses to show percentages of questions and reflections used during the interaction. The task proposed in our work is related to behavioral coding, but our focus is on detecting the overall quality of a specific verbal behavior (a reflection), rather than identifying their type.

Our scoring system can also be used as a component in larger systems for writing or rewriting counselor utterances, similar to the framework in (Laban et al., 2020; Sharma et al., 2021). Finally, Pérez-Rosas et al. (2019) show that high reflection frequency is associated with high-quality, demonstrating the importance of practicing reflective behavior in MI.

3 Motivational Interviewing Reflection Dataset

To build our reflection scoring models, we compile a dataset consisting of brief interactions between counselors and clients portraying different levels of reflective listening skills. Each interaction is in English and includes a client prompt, i.e., a client's statement that is usually given to the counseling trainee, paired with counseling responses portraying different levels of reflections skill, i.e., low quality, medium quality, and high quality. We build the dataset using both expert and crowd-sourced annotators and also leverage conversational data from an MI dataset to obtain additional prompt-response pairs from conversations snippets containing reflections. We make our data available at `https://lit.eecs.umich.edu/downloads.html#PAIR`.

3.1 Annotations

We classify reflections into Simple Reflections and Complex Reflections based on standard criteria (McMaster and Resnicow, 2015; Moyers et al., 2016). We also consider a third category of responses consisting of Non Reflections.

Simple Reflection (SR). This entails reflecting back to the client on what they said, using different words (paraphrasing). Simple reflections typically do not include new insights or inferences. They tend to capture what was just said more than what lies behind or ahead of the client statement. We categorize SRs as mid-quality reflections, whose quality lies between that of complex and non-reflections. In Table 1, the response "You believe you will die from breast cancer, just like your mom." is a medium quality reflection containing a simple reflection because it adds no additional meaning to what the client has already expressed.

Complex Reflection (CR). Complex reflections entail the counselor adding or inferring something new from the client statement. This may include naming a feeling or emotion that has not yet been expressed by the client; inferring why the client might have said something; or stating where they are headed. As an example, the counselor utterance "Your mother's death was devastating. You're worried you may die the same way she did." in Table 1 is a complex reflection (i.e., high quality response)

Source	Prompt	Response	Label
Expert Annotated	My mother died of breast cancer, so I know I'm going to die of it too.	Your mother's death was devastating. You're worried you may die the same way she did.	CR
		You believe you will die from breast cancer, just like your mom.	SR
Crowdsourced		Are you giving up?	NR
		There are measures you can take to prevent that. Remember, you are not your mother.	NR
MITI Sessions	At the computer. I like to play games on the computer. So ... Then I smoke then. Watching TV.	So it really is integrated. It's really a big part of your life.	CR

Table 1: Example responses and their reflection labels from collected datasets

as it brings attention to the client's traumatic experience, rather than merely rephrasing what was said. Complex reflections are considered a high quality MI response.

Non Reflection (NR). Responses that include unsolicited advice or asking questions when a reflection would have been a better response are classified as low quality responses.

3.2 Prompt-reflection Pairs Using Hand-crafted Prompts

We start by collecting reflections for a set of 318 manually crafted client prompts, covering a wide range of health related behaviors such as diabetes, weight management, smoking cessation, vaccination, or alcohol consumption. The client prompts are sourced from MI training materials used during a graduate MI class taught by one of the paper authors. We conduct our annotation process to create responses to each client prompt with varying reflection quality levels.

Expert Annotations. We recruit two annotators with MI expertise to write high, medium and low quality reflections (CRs, SRs, NRs, respectively) for each of the prompts using the definitions for each type of reflection. We assign half of the prompts to each annotator and ask them to write two complex reflections, one simple reflection and two non-reflections for each client prompt.

Non-expert Annotations. One concern with only using expert annotated data is that experts might struggle to simulate MI inconsistent responses, which we want to capture in low quality (NR) responses. To tackle this problem, we leverage crowdsourced annotations from non-MI experts, using workers from Amazon Mechanical Turk. Our rationale for using non-expert annotation for low quality (NR) reflections is that non-experts responses should be closer to inexpert MI practitioner responses that we expect our models to encounter during training or evaluation scenarios. Moreover, this strategy allow us to generate low quality reflections without the need of expert input. During the annotation process, we request AMT workers to provide "advice" to the situation described in the client prompt so their responses will likely contain directive language that does not follow MI guidelines. This step is inspired by expert observation that providing unsolicited advice is a common behavior while trainees are still learning to craft reflections. To improve the diversity of the responses, we request three responses per prompt, and each response is annotated by a unique AMT worker. We manually verify the responses and reject those that fail to follow the provided guidelines.

Our final set of prompt-reflection pairs using the hand-crafted prompts consists of two complex reflections, a simple reflection, and five non reflections for each of our 318 client prompts. This results in 2,544 prompt-response pairwise examples. Table 2 shows the average number of tokens for each type of reflection in the dataset.

3.3 Prompt-reflection Pairs from Counseling Conversations

We also obtain prompt-reflection pairs from an existing counseling dataset by Pérez-Rosas et al. (2016) containing annotations for 2690 simple reflections and 2876 complex reflections. We use the dataset annotations to select conversation snippets leading to counseling reflections.

To build client prompt-reflection pairs, we take the previous utterance by the client as the prompt and collect responses that are labeled as complex and simple reflections.The statistics of the resulting dataset are shown in Table 2. We use this dataset for additional validation and sampling of non-matching responses for our learning objectives

Dataset	Prompts	Average number of tokens					
		All	Prompt	CR	SR	NR-Experts	NR-Crowd
Hand-crafted prompts	318	27	48	31	14	20	26
MI Conversations	4365	31	31	33	27	NA	NA

Table 2: Dataset statistics for each source. "NR-Crowd" standards for low quality reflections collected from crowdsource workers.

in Section 4.

4 Prompt-Aware margIn Ranking (PAIR)

Our reflection scoring task consists of assigning a reflection quality score s between [0, 1] to a interaction pair containing a client prompt p and a candidate reflection by a counselor r. While this task can be viewed as regression, obtaining ground truth labels is expensive and noisy, even for expert annotators. Instead, we formulate the scoring problem as a learning-to-rank task, in which the training data labels are pairwise relevance levels based on the ground truth reflection, i.e, NR, SR and CR (Cao et al., 2007).

We develop a scoring framework inspired by contrastive and metric learning strategies, where binary contrastive estimations are computed between examples for consecutive reflection quality levels. Our model combines two metric-learning based objectives that enforce the correct ranking along with prompt relevance. Figure 2 shows an overview of our training process.

For our model backbone, we adopt a transformer-based encoder architecture (Vaswani et al., 2017), but our learning objectives can be extended to other neural models, such as recurrent neural networks. We use the RoBERTa variant of the transformer-based encoder (Liu et al., 2019). We implement a simple cross encoder that takes the concatenated sequence of a prompt and a response pair as input. Since this design choice allows to directly model the interaction of prompt and response tokens, we classify our main model as a cross encoder-based model, following the characterization of encoder provided by Humeau et al. (2020).

Multi-level Margin Ranking Objective. We introduce a margin ranking loss term to ensure a distance gap between quality levels of reflections taking inspiration from Lin et al. (2020). The ranking objective uses a margin parameter μ or 2μ, depending on the examples being compared in the loss term, where μ corresponds to the "gap" in scores between each neighboring quality level pair. Next, we use μ when the quality gap is within one level i.e, distinguishing between medium quality and high quality pairs (SR, CR) or low quality and medium quality pairs (NR, SR), and 2μ when the gap is within two levels i.e, low quality and high quality pairs (NR,CR). The loss is calculated using the equation below, where p is the client prompt and r_{CR}, r_{SR}, r_{NR} respectively denote CR, SR, and NR responses to p.

$$\begin{aligned}\mathcal{L}_{\text{gap}} = &\max\{0, \mu - (s(p, r_{CR}) - s(p, r_{SR}))\} \\ &+ \max\{0, \mu - (s(p, r_{SR}) - s(p, r_{NR}))\} \\ &+ \max\{0, 2 * \mu - (s(p, r_{CR}) - s(p, r_{NR}))\}\end{aligned}$$

Prompt-Aware Margin Ranking Objective. In preliminary experiments using $\mathcal{L}_{gap}$ we find that the model tended to ignore the client prompt when making predictions. This can result in incorrect scoring for cases where responses are not related to the client prompt but do follow MI style. To avoid this problem, we design a prompt-aware objective to penalize the model against this scenario.

To provide examples of such cases to the model, we build an additional set of pairs by sampling CR and SR responses (m_{CR}, m_{SR}) from the training batch and matching them with random prompts from the same batch (p), with the condition that the matched prompts must be different from the original pairs. Then, we treat the constructed pairs of prompt and mismatched responses as low-quality examples (NR). We thus formulate the following prompt-aware ranking objective, where r_{CR}, r_{SR}, μ, and 2μ are defined as in $\mathcal{L}_{gap}$, while m_{CR}, m_{SR} refer to the mismatched responses.

$$\begin{aligned}\mathcal{L}_{\text{prompt}} = &\max\{0, 2 * \mu - (s(p, r_{CR}) - s(p, m_{CR}))\} \\ &+ \max\{0, \mu - (s(p, r_{SR}) - s(p, m_{SR}))\}\end{aligned}$$

Our scoring function is the transformer encoder model, followed by a pooling layer and a sigmoid activation. For our final model, we combine the $\mathcal{L}_{gap}$ and $\mathcal{L}_{prompt}$ objectives with equal weights:

$$\mathcal{L} = \mathcal{L}_{gap} + \mathcal{L}_{prompt}$$

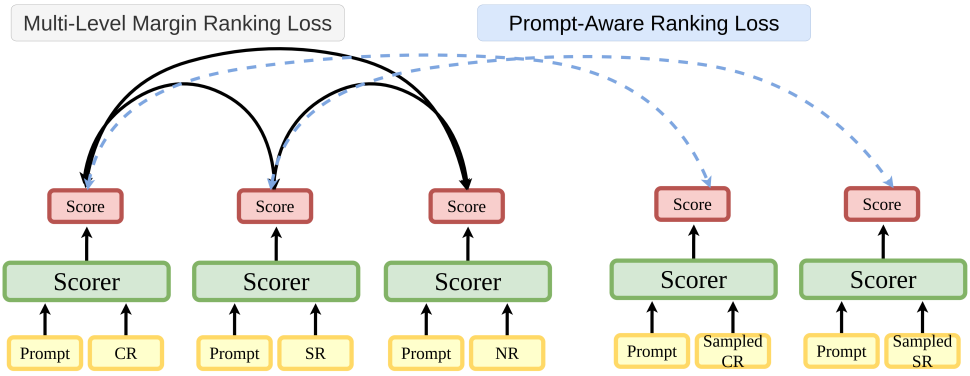

Figure 2: Diagram of our model training process. The black arrows represent the comparison of varying levels of reflections within a prompt-response tuple for the multi-level margin ranking loss. The dashed blue arrows represent the comparison of simple and complex reflections against sampled reflections from a different prompt-response sample.

5 Experiments

5.1 Experimental Setup

For all our models, including baselines, we use the Roberta architecture from (Liu et al., 2019) and initialize our models with pretrained weights `mental-roberta-base` from (Ji et al., 2022). Our choice of pretrained weights is motivated by our domain being similar to that of the pretraining corpus used for `mental-roberta-base`, which contains mental-health topic posts from Reddit, in which counsel-seeking posts are paired with responding comments. Additionally, we conduct preliminary experiments using the pretrained weights and find that they improve overall performance.

We implement our models using the PyTorch (Paszke et al., 2019) and Huggingface Transformers (Wolf et al., 2020) packages. For training, we use the Adam optimizer with weight decay of 0.01, a constant learning rate of $2e^{-5}$, and a batch size of 64 samples. We also apply a dropout rate of 0.1 to all layers. To fit the training data into our computing device in an efficient manner, we subsample each data row into a smaller row. That is, given a prompt-tuple with one prompt and eight responses (2/1/5 CR/SR/NR), we generate 20 sub-tuples with one prompt and four responses, composed of 1 CR, 1 SR and 2 NR. In this manner, the total number of pairwise data is 318 * 20 = 6360. We train for two epochs on one NVIDIA GeForce RTX 2080 Ti, with a batch size of 64 (using gradient accumulation).

For the evaluation of our models, we set aside 20% of our data as our test set. As our main performance metrics, we use recall@1, Pearson and Spearman, and Kendall's Tau correlation coeficient. We compute the Pearson and Spearman correlations between the model-predicted scores and the discrete label mapped to an integer level corresponding to their order. For recall@1 and Kendall's Tau, given a client prompt, counselor responses with varying levels of reflection quality are considered rankings with respect to true and predicted reflection levels, and we use the true and predicted rankings to compute the mentioned metrics.

Further, to test the model performance on prompt and response pairs that do not match (i.e., the response is not a coherent reply to the prompt), we augment each prompt-response tuple with a k response sampled from a different prompt tuple. The ground truth judgement for the randomly matched responses are NR, regardless of the original judgement the responses received.

5.2 Reflection Scorer Models

Using the PAIR model as our architecture, we experiment with different versions of the model by combining the multi-level ranking and prompt-aware objectives. Specifically, we compare models with or without the prompt-aware objective to test the effectiveness of our proposed method.

5.3 Baselines

We compare our model with a set of baselines that share the same transformer encoder architecture and pretrained weights. Additionally, we experiment with a classifier and a regressor using linear heads on top of the encoders.

Naive Classifier. Given a prompt and a response, it outputs a discrete label for the reflection quality

Metrics / Model	Naive Classifier*	Naive Classifier	Naive Regressor*	Naive Regressor	PAIR*	PAIR
Recall@1	0.8952	0.8349	0.9174	0.5873	**0.9253**	0.6444
Pearson	0.8713	0.7652	**0.8994**	0.7998	0.8722	0.7205
Spearman	**0.8816**	0.7858	0.8784	0.7994	0.8811	0.7415
Kendall's Tau	0.6955	0.5685	0.8653	0.7389	**0.8694**	0.7216

Table 3: Evaluation results on the set-aside test set. Our final model is PAIR. For Pearson and Spearman correlations, the values are statistically significant with p-value < 0.05.

Metrics / Model	Naive Classifier*	Naive Classifier	Naive Regressor*	Naive Regressor	PAIR*	PAIR
Recall@1	0.8952	0.8349	0.9174	0.5873	**0.9253**	0.6444
Pearson	0.4892	0.6868	0.5317	0.6902	0.5108	**0.7396**
Spearman	0.4896	**0.7227**	0.5018	0.6590	0.5001	0.6795
Kendall's Tau	0.2397	0.4316	0.4539	0.5824	0.4485	**0.5940**

Table 4: Evaluation results on the set-aside test set augmented with randomly-matched responses. Our final model is PAIR. For Pearson and Spearman correlations, the values are statistically significant with p-value < 0.05.

of the responses i.e., NR, CR or NR. The classification model is trained using standard cross entropy loss against a set of discrete reflection quality labels included in our annotated dataset.

Naive Regressor. Given a prompt and a response, it outputs a scalar score (between [0,1]) as the reflection quality level of the response. This model is trained using standard mean squared error loss. To train this model, we convert discrete labels into continuous scores using the following mapping: {CR: 1.0, SR: 0.5, NR: 0.0}.

For a fair comparison, we also consider versions of the baselines that are additionally trained on a prompt-aware loss term on top of original losses. As in our cross-encoder model, we introduce prompt-aware negative examples by switching the client context and labeling it as NR.

6 Results

Table 3 shows the evaluation results of our scorer models and baselines on the set-aside test set, while Table 4 shows the experiment results over the test set augmented with randomly-matched responses. In both tables, PAIR refers to our main models trained with the full set of our objectives, while * indicate that we remove the prompt-aware objective for ablation, and only leave the multi-level margin ranking objective during training.

For Tables 3 and 4, the recall@1 results are identical, indicating that even after randomly matched responses are added, all the models are able to correctly identify response with the highest reflection level (complex reflection). Moreover, we note that for the naive classifier models Spearman correlation scores higher than Pearson correlation, likely due to the fact that Spearman correlation measures monotonic relationships and the naive classifier predictions contain frequent ties, since unlike other models the classifier outputs discrete integers.

Baseline Comparison. We compare the performance of our model against several baselines. When the models are tested against data without randomly matched responses (Table 3), the best performing baseline models (Naive Classifier, Naive Regressor) perform similarly to PAIR*, which uses the multi-level objective but not the prompt-aware objective. Although the PAIR* model scores highest of recall@1 and Kendall's Tau, its Pearson and Spearman correlation coefficients are slightly worse than the naive models.

However, in Table 4, we see more evidence in favor of PAIR. Comparing the best performing models (the prompt-aware models), we see that the PAIR model, which uses both the multi-level and prompt-aware objectives perform better than the Naive Regressor model. When comparing the Naive Classifier and the PAIR model, we note that the Naive Classifier models are better for the recall@1 metric. We remark that the two models are not directly comparable, since they represent different frameworks of prediction and feedback. However, we argue that in a setting where a continuous score is desired, our model is preferable to the classifier, since it can provide a more detailed feedback, which can better convey the implicit preference ranking of different responses, as evidenced by its higher Kendall's Tau score.

7 Ablation Study

We study the effects of the prompt-aware learning objective by conducting a set of ablation experiments. Specifically, we evaluate our scorer models and baselines with or without the prompt-aware learning objective. In Table 3, when we tested our models on test cases where all prompt-response pairs were matched pairs (i.e. the response was in response to the matched prompt), prompt-aware models perform worse in all metrics than their * counterparts, showing that using the prompt-aware during training leads to performance losses when tested on data without randomly matched responses. This indicates that there is a performance trade-off between reflection scoring and incoherence detection.

When we test our models on a dataset augmented with randomly-matched negative responses, we find that the prompt-aware loss leads to improved performance on data that includes random responses. In Table 4, prompt-aware models perform consistently better than their counterparts on all metrics except recall@1. This result is expected, as the prompt-aware loss was specifically designed to prevent the model from ignoring prompts when predicting reflection levels. We argue that our objective is effective in addressing this problem, which can be a critical point of failure, especially when the assumption of coherent or relevant user input might not be guaranteed.

8 User Study

In addition to testing our model on annotated data, we also deploy our final model (PAIR) in a real-life education setting – a graduate-level MI training course taught by one of the authors of this paper. We collaborated with MI experts at the University of Michigan School of Public Health and implemented a web application for a graduate-level MI training course[1]. We develop and evaluate a web based application that uses the PAIR model to provide real-time scoring feedback to students while learning to create reflective responses to a given client prompt. We plan to make the system website available for demonstration.

System Implementation. The web platform is text-based and shows a client prompt to the counseling trainee and ask them to write a reflective response for the situation depicted in the prompt. The system shows five prompts at a time, but the trainee only needs to provide at least one response to receive feedback. After the trainee has provided their response(s), the system shows detailed feedback for each response(s) consisting of a numerical score ranging between 0-1 and two examples of high quality (CR) reflections for the given prompt.

The system is implemented as a web server using Nginx,[2] Gunicorn,[3] and the Flask[4] web framework and is run on a secure machine. The system takes less than one second to run the PAIR model for $\tilde{3}0$ prompt and response pairs and provide feedback.

Participants. We conduct a user study with 30 students enrolled in the MI training class. The students used our system to complete three assignments that required them to practice their reflective skills. Over the course of four weeks between January - February 2022, they completed three assignments, each consisting of a set of client prompts designed by the course instructor. Before using the system, participants were directed to a page where they read the consent form. If they agreed to participate, they were directed to the main system view showing the different prompts to be answered for the given assignment.

For each assignment, the participant was presented with about 20 client prompts, to which they were asked to write a reflective response to the situation being described by the client. Participants completed their assignments at their own pace as the system allowed them to save and retrieve their work at any time. After the participant submitted their responses, the web application ran our model in the server and provided detailed feedback to each response, consisting of two ground truth high quality reflections for each prompt, and the model predicted score. Participants were evaluated on the basis on completion/non completion of the given assignment. After completing each assignment, participants took a survey regarding the system accuracy and usability and optional qualitative feedback. A screenshot of our web interface can be found in Figure 4.

8.1 Evaluation with User-generated Responses

The student submissions are new data that can be useful to further train or evaluate our models. We

[1] https://sph.umich.edu/academics/courses/syllabi/HBEHED671.pdf

[2] https://www.nginx.com

[3] https://gunicorn.org

[4] https://flask.palletsprojects.com/en/2.1.x/

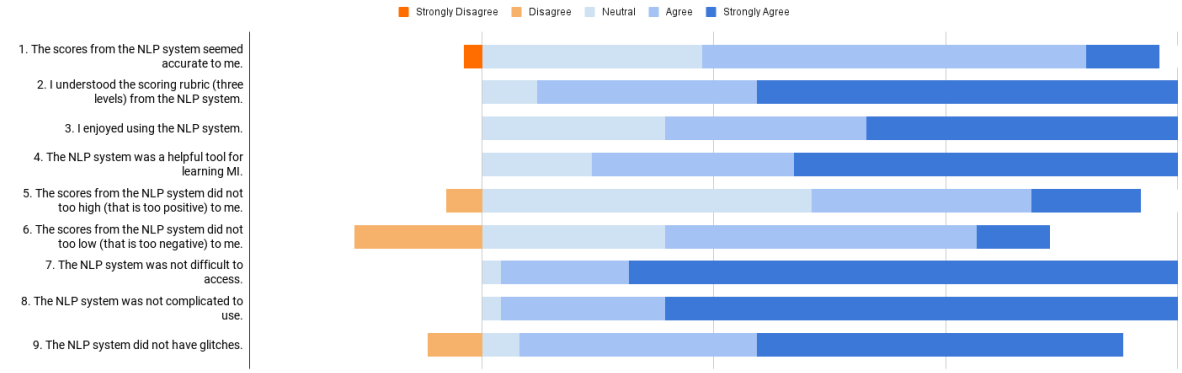

Figure 3: User survey results on a 5-point Likert scale. For comparing answers in a unified positive scale, questions 5-9 were negated.

*-3mm

GT / Model	CR	SR	NR	Accuracy
CR	2341	183	14	0.9223
SR	32	324	26	0.8481
NR	1	24	54	0.6835
Pearson Correlation: 0.7829				
Spearman Correlation: 0.5776				

Table 5: Confusion matrix of the model predictions on student submissions.

annotate the submissions with the help of a student instructor, who reviewed each submitted response alongside the original prompt and annotated it with a discrete reflection level judgement. Given this new annotated set, we evaluate the performance of our model using accuracy and correlation metrics, as shown in Table 5. Specifically, we convert a model predicted score into a discrete judgement using the mapping: $\{\text{CR} : [0.7, 1.0], \text{SR} : [0.3, 0.7], \text{NR} : [0.0, 0.3]\}$. This mapping is based on our and experts' observation on score distribution over different levels of responses.

As indicated by the confusion matrix and accuracies in Table 5, our model does best on correctly identifying CRs, while performing less well on SRs and NRs. As the distribution of ground truth labels shows, identifying and encouraging reflective listening is a priority of this class, and hence the low false positive rate shown by the system is aligned with this design objective.

8.2 Usability Evaluation

We also conduct a usability study to test how well our model does in a real setting.

Usability Survey Result. To collect and measure user experience and satisfaction, we devise a 5-point Likert questionnaire on the perceived accuracy and usability of our system. Figure 3 shows the 9 questions covering model error and system usability, and the distribution of user responses. Overall, our system received positive assessment on average for both accessibility and performance questions.

Qualitative Feedback. We also asked users to submit free-form text feedbacks after submitting the assignment. Among the submitted comments, positive answers focus on how the application allowed them to have more practice and build up confidence, while negative feedback is usually concerned with the functionality aspects such as saving and loading their work.

9 Conclusion and Future Work

In this work, we introduced the task of reflection scoring and developed a prompt-aware margin ranking approach, PAIR, to tackle this problem. Our model learns to predict continuous scores from discrete label training data and outperforms simple baselines on several metrics, and we showed its deployment in an educational setting with real students and instructors. We plan to extend our model to incorporate diverse information that can assist counselors in understanding their clients, such as dialog context, client background, or medical knowledge.

Acknowledgements

The authors would like to express gratitude toward researchers and students from the University of Michigan School of Public Health, for their valuable feedback and participation in this project. This material is based in part upon work supported by the Precision Health initiative at the University of

Michigan and by the John Templeton Foundation (grant #61156). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Precision Health initiative or the John Templeton Foundation.

10 Ethics Statement

Privacy and Data Protection. We ensure that users of our systems are informed of our data collection practices. Moreover, we conduct data cleaning and anonymization to remove any personal or sensitive information from the collected data.

Bias and Impact of the Model Since our model provides feedback to human behavior, there is a risk that the model may have negative consequences. For instance, biases or artifacts contained in expert annotation can be encoded in such models and may exert influence on students who are trying to mimic or learn from the model. Although we have not detected any such examples or trends during the model testing and deployment, we plan to further study and evaluate the impact of our models as future work.

11 Limitations

Our work has several limitations, which we aim to address in our future work.

First, our model is finetuned using only synthetic data, annotated by a group of experts, for a predefined collection of simulated client prompts. We included real counseling data in our framework through pretraining, but this data is not directly used in the supervised training or downstream evaluation of the model. Although we evaluate our model in the wild through system deployment and user evaluation, we hope to further understand and bridge the gap between models trained using observed data and models trained using synthetic data.

Second, our models rely on linguistic and reflection style to score reflections but do not account for conversational aspects such as empathy, which is also considered an important counseling strategy while generating reflections. In one of our preliminary experiments where we compared our reflection scores and empathy levels computed by Sharma et al's empathy model (Sharma et al., 2020), we observed that emotional reactions to mental health posts tend to have low reflection levels, indicating that counseling reflection is related but not identical to empathy in counseling.

Finally, the reflection scoring system proposed in this paper mainly provides numerical scoring feedback to trainees along with good reflection feedback that has been designed by the course instructor. We are working on expanding the system to include models for different types of feedback, beyond mere reflection level scoring. For instance, by exploring generative models to automatically create counselor responses, reference responses can be provided for students, even when annotated ground truth is unavailable. Additionally, rewriting models can provide more valuable feedback by presenting improved versions of students' own responses.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*.

Victor Ardulov, Torrey A. Creed, David C. Atkins, and Shrikanth Narayanan. 2022. Local dynamic mode of cognitive behavioral therapy.

Norma G. Bartholomew, George W. Joe, Grace A. Rowan-Szal, and D. Dwayne Simpson. 2007. Counselor assessments of training and adoption barriers. *Journal of substance abuse treatment*, 33 2:193–9.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. volume 227, pages 129–136.

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. *CoRR*, abs/2102.11573.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *ICLR*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.

Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association*

for Computational Linguistics, pages 5135–5150, Online. Association for Computational Linguistics.

Zibo Lin, Deng Cai, Yan Wang, Xiaojiang Liu, Haitao Zheng, and Shuming Shi. 2020. The world is not binary: Learning to rank with grayscale data for dialogue response selection. In *EMNLP*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Fiona McMaster and Kenneth Resnicow. 2015. Validation of the one pass measure for motivational interviewing competence. *Patient education and counseling*, 98 4:499–505.

Theresa Moyers, Lauren Rowell, Jennifer Manuel, Denise Ernst, and Jon Houck. 2016. The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017a. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435, Vancouver, Canada. Association for Computational Linguistics.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J. Goggin, and Delwyn Catley. 2017b. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137, Valencia, Spain. Association for Computational Linguistics.

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.

Ashish Sharma, Inna Wanyin Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. *Proceedings of the Web Conference 2021*.

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

A Additional Information of Our User Study

A.1 Obtaining Consent from Participants

Before they could access the assignments, participants were asked to read and sign an informed consent form, which informs that their submissions will be securely stored and be used for academic research. In the case that some participants were not comfortable doing this assignment, they could choose from alternatives provided by the class instructor, but no one opted to do so in this study.

A.2 Data Anonymization and Protection

To ensure that user data is securely stored without compromising privacy, we only ask 8-digit student IDs for assignment submission, which then are mapped to unrelated hash strings for storage in a secure server.

A.3 Web Interface

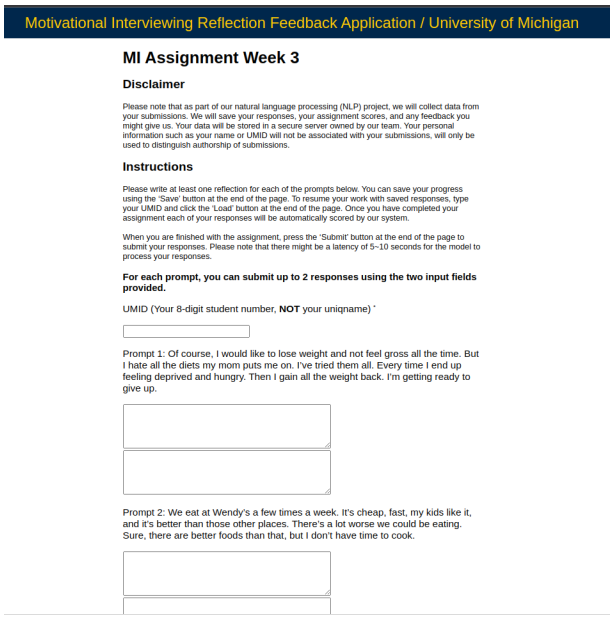

Figure 4: A view of our web interface