Evaluating Language Models for Assessing Counselor Reflections

DO JUNE MIN, Department of Electrical Engineering and Computer Science, University of Michigan, USA
VERÓNICA PÉREZ-ROSAS, Department of Electrical Engineering and Computer Science, University of Michigan, USA
KENNETH RESNICOW, School of Public Health, University of Michigan, USA
RADA MIHALCEA, Department of Electrical Engineering and Computer Science, University of Michigan, USA

Reflective listening is a fundamental communication skill in behavioral health counseling. It enables counselors to demonstrate an understanding of and empathy for clients' experiences and concerns. Training to acquire and refine reflective listening skills is essential for counseling proficiency. Yet, it faces significant barriers, notably the need for specialized and timely feedback to improve counseling skills. In this work, we evaluate and compare several computational models, including transformer-based architectures, for their ability to assess the quality of counselors' reflective listening skills. We explore a spectrum of neural-based models, ranging from compact, specialized RoBERTa models to advanced large-scale language models such as Flan, Mistral, and GPT-3.5, to score psychotherapy reflections. We introduce a psychotherapy dataset that encompasses three basic levels of reflective listening skills. Through comparative experiments, we show that a finetuned small RoBERTa model with a custom learning objective (Prompt-Aware margIn Ranking (PAIR)) effectively provides constructive feedback to counselors in training. This study also highlights the potential of machine learning in enhancing the training process for motivational interviewing (MI) by offering scalable and effective feedback alternatives for counseling training.

CCS Concepts: • **Computing methodologies → Natural language processing**; • **Applied computing → Health care information systems**; • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Motivational Interviewing, Computational Counseling, Reflective Listening, Large Language Modeling

ACM Reference Format:
Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2018. Evaluating Language Models for Assessing Counselor Reflections. In *ACM Transactions on Computing for Healthcare Special Issue on Large Language Models, Conversational Systems, and Generative AI in Health.* ACM, New York, NY, USA, 24 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Counselor training is expensive and time-consuming due to the extensive expert supervision involved [4]. Current strategies for counselor training usually rely on either role-playing or monitoring and live recording video interactions, which are then manually evaluated to provide constructive feedback, thus limiting counselors' opportunities to practice and receive timely evaluative feedback.

While several promising approaches have been proposed to automatically provide evaluative feedback to counselors [10, 54, 56, 59], generating helpful feedback in real-time remains a challenge. This is particularly the case in

Manuscript submitted to ACM

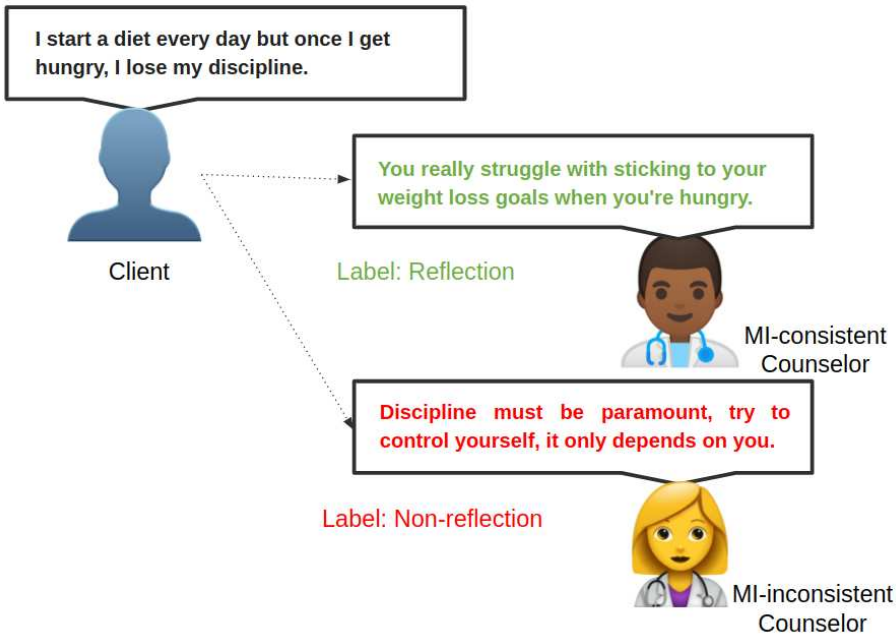


Fig. 1. Examples of Reflective and Non-reflective Counselor Behaviors.

educational settings, where counseling trainees could benefit from supportive learning environments that allow them to make mistakes and learn at their own pace while acquiring counseling skills.

Seeking to address this need, we study the task of quantitatively evaluating the language of counseling trainees when learning to formulate responses to clients' statements. We believe that the automatic assessment of counselors' verbal behavior can enhance their training by allowing them to practice reflective listening skills in real-time and provide immediate feedback. Among core counseling skills, we focus on responses containing reflections, i.e., counselor statements aiming to understand and reflect on what the client is saying. Figure 1 shows an example of a counselor's reflection in response to a client's situation.

We experiment with two approaches for automatic reflection assessment. First, fine-tuning a small transformer model using a novel margin ranking-based approach that can output a continuous score learned from discrete annotations of counseling reflections (PAIR (Prompt-Aware margIn Ranking). Second, we use large language models (LLM) and in-context learning to obtain reflection-quality evaluations.

We conduct a set of comparative experiments to evaluate each system's ability to learn the correct ranking of counseling responses. Additionally, we evaluate a real scenario in which we deployed our fine-tuned system (PAIR) in an educational setting. We conducted 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) Two LM-based frameworks for reflection scoring using contrastive learning approaches (PAIR, Prompt-Aware margIn Ranking), and LLMs with in-context learning; and (3) Quantitative and qualitative assessments of our models on the annotated dataset and through in-the-wild deployment and feedback.

While not addressed directly in this work, having the ability to evaluate the quality of a counseling response can serve as a guiding signal for dialog-based systems to provide evaluative feedback during the acquisition of counseling skills. This could be implemented using text style transfer and controlled generation approaches, which have been successfully used in the past for increasing politeness or empathetic tone in user responses [28, 35, 37].

.

Hence, an important application of our scoring system is being the evaluation component for a counselor response rewriting system that provides trainees with direct suggestions for improving their responses. Furthermore, our scoring system could also be integrated into a generation system to guide the production of responses that are more closely aligned with the principles of effective counseling. We thus believe that our work aligns with the broader objectives of utilizing NLP to enhance the training and proficiency of care providers, as evidenced by recent studies on the application of LLMs in the mental health domain [6, 10–13, 57, 58, 67].

2 RELATED WORK

NLP and Behavioral Counseling. Automated analysis and evaluation of verbal strategies used in mental health conversations has emerged as a promising intersection of psychotherapy and NLP [1]. With rising awareness of the increased need for mental health care, several NLP models or techniques that aim to enable scalable, efficient processing and analysis of counseling language have been proposed to understand counseling interactions[21, 39, 43]. Work has also been done on addressing evaluation and feedback in counseling by measuring the fidelity to treatment via automatic behavioral coding [2, 14, 47]. Work on this includes predicting and forecasting counselor behaviors, including questions, reflections, or change talk [8], and also evaluating conversational aspects such as empathy, verbal mimicry, and conversational tendencies [33, 46, 53, 66]. More recently, dialog-based systems have been explored to address the issue of mental workforce shortage and assist in developing and evaluating basic counseling skills. Tanana et al. [59] developed a patient-like conversational agent that interacts with counselors while practicing open questions and reflections. It categorizes their responses to show percentages of questions and reflections used during the interaction. Shen et al. [56] generated responses containing reflections using LLMs and context expansion strategies using retrieval of relevant responses from previous interactions and expanding keywords from the client utterances. Subsequent work [55] explored the inclusion of domain-specific and medical knowledge to be integrated into the generation of reflective responses.

Our work proposes a task related to behavioral coding. However, we focus on detecting the overall quality of a specific verbal behavior (a reflection) rather than categorizing it.

Contrastive and Metric Learning. Contrastive learning focuses on learning representations by contrasting positive pairs against negative pairs, effectively teaching the model to distinguish between closely related examples [24]. Metric learning extends this concept by aiming to learn a distance function that can measure the similarity or dissimilarity between pairs in a meaningful space [25]. Our work uses contrastive learning to frame the scoring problem as a learning-to-rank issue where training data labels denote pairwise relevance levels based on reflection quality [9, 30]. Inspired by works such as Lin et al. [30], we use binary contrastive estimations between examples of consecutive reflection quality levels to refine model training. Although initial experiments considered contrastive representation learning approaches like those proposed by Gao et al. [15], Liu et al. [31], we opt for using margin-ranking objectives combined with a cross-encoder architecture for both prompt and response analysis, as they showed more adaptability for our specific application.

Prompt	Response	Quality	Source	Definition
My mother died of breast cancer, so I know I'm going to die of it too	Your mother death was devastating. You are worried you may die the same way she did.	Complex Reflection (CR)	Expert	A complex reflection adds meaning or emphasis, moving beyond what the person said to infer deeper concerns.
	You believe you will die from breast cancer, just like your mom	Simple Reflection (SR)	Expert	A simple reflection stays close to what the person said, simply restating or paraphrasing their words.
	You need to have genetic testing in order to know your own personal risk. We cannot make clinical judgments based on your mom.	Non-Reflection (NR)	Crowdsourced	A non-reflection offers information or advice without reflecting the speaker's emotions or thoughts.

Table 1. Example prompt-response pairs and their reflection labels

Large Language Models (LLMs). Recent research and application trends in machine learning have embraced using LLMs. These models are usually transformer-based generative models with over several billion parameters requiring extensive training on commercial-scale computing infrastructure. While more challenging to train and run on most personal computing hardware, these models boast state-of-the-art performance on many natural language benchmarks and offer easy adaptability to a wide array of domains, including mental health and psychotherapy [6, 10–13, 57, 58, 67]. Several concerns regarding the deployment of LLMs as opaque systems potentially exacerbating implicit biases and stereotypes or causing unintended detrimental outcomes have prompted practitioners to exercise caution when integrating LLMs into patient-facing therapeutic applications [26, 29]. Concurrently, there has been a mounting interest in leveraging LLMs to enhance care providers' and clinicians' training and proficiency. These applications hold promise for enhancing the capabilities of care providers and clinicians. By harnessing the power of LLMs, professionals can access a wealth of data-driven insights and personalized recommendations, thereby improving diagnostic accuracy and treatment efficacy [19, 52]. In this regard, our work focuses on the effectiveness of NLP in providing scalable and precise feedback for improving reflective listening skills in motivational interviewing training.

A particularly promising aspect of LLMs in therapeutic applications is their capability for in-context learning (ICL) and prompting [18, 34]. ICL allows LLMs to generate responses or perform tasks relevant to the provided context or instructions without the need for additional fine-tuning [38]. This is especially advantageous in mental health, where understanding and generating nuanced language can significantly enhance therapeutic interactions [64]. While prompting LLMs with specific scenarios or questions related to mental health can guide the models to apply their generalized knowledge in ways that are directly beneficial to counseling and therapy, the design of effective prompts for LLMs in mental health applications requires careful consideration of the therapeutic context, the objectives of the interaction, and the client's specific needs [27, 65]. This approach not only capitalizes on the linguistic capability of LLMs but also directs their capabilities toward supporting mental health professionals in providing high-quality care.

3 REFLECTIVE LISTENING DATASET

Motivational interviewing (MI) is a counseling style that motivates clients to make behavioral changes through collaborative conversation. MI counselors are expected to use standard core counseling skills when engaging with their clients. Among them, reflective listening is one of the most critical skills counselors must develop to become proficient in motivational interviewing (MI). It entails responding in a way that recognizes and delves into the significance of what the client has shared during the conversation [3]. Previous studies have shown that the quality and quantity of reflection in counselor behavior is empirically correlated with the perceived quality of counseling [47] and treatment outcome [16]. Given the importance of acquiring reflective listening skills and the need for actionable and immediate

feedback during this process, our work focuses on automatically providing evaluative feedback to counseling trainees in real time.

3.1 Conversational Prompts

We compiled a new dataset of brief interactions between counselors and clients portraying different levels of reflective listening skills. Each interaction is in English and includes a conversational prompt with a counseling scenario that likely leads to a reflective response—usually given to the counseling trainee when learning to elicit reflective responses, see an example in Table 1. For the remainder of the paper, we refer to these as client prompts.

We build the dataset using both expert and crowd-sourced annotators and leverage conversational data from an existing counseling dataset [45] annotated with reflections to obtain additional prompt-response pairs containing reflections.

Hand-crafted Prompts. We manually crafted 318 prompts with the assistance of a Motivational interviewing expert, who is also one of the authors of this paper. The prompts cover health-related behaviors such as diabetes, weight management, smoking cessation, vaccination, and alcohol consumption. We use these prompts to collect responses from expert and non-expert annotators to portray diverse reflection skills.

Prompts from Counseling Conversations. We also use data from an existing conversational counseling dataset [45]. The dataset contains MI counseling conversations with MITI annotations for counselor utterances. We use the reflection annotation subset to extract prompt-reflection pairs by taking the previous client's utterance as the prompt along with counselor responses labeled as complex and simple reflections. We thus obtained 4,365 client prompt-counselor reflection pairs, including 2,429 prompt-CR and 1,636 prompt-SR pairs. The statistics of the resulting dataset in terms of the average number of tokens per prompt and reflection quality type, are shown in Table 2. Since this dataset is lower in quality than our hand-crafted set due to annotation quality and style difference (spoken vs written), we only use this data for validation and user study.

3.2 Expert Annotations

Two psychotherapists with MI expertise annotate the hand-crafted prompts. We ask them to write complex, simple, and not-reflection responses for a given prompt using the guidelines of the Motivational Interviewing Treatment Integrity (MITI) [40] scheme, the current gold standard for assessing the integrity of Motivational Interviewing interventions.[1] Annotators had previously undergone MITI training and had worked together on similar annotation tasks. However, they worked independently, and each annotated half of the available prompts.

We use the MITI definitions as guidelines for labeling simple and complex reflections. Additionally, we defined a third category to label responses that showed poor or nonexistent reflective skills.

Simple Reflection (SR). These responses reflect what the client said, using different words, e.g., 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's statement. 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. We categorize SRs as mid-quality reflections, whose quality lies between complex and non-reflections.

[1] MITI https://casaa.unm.edu/assets/docs/miti1.pdf

Dataset	#Prompts	Average number of tokens					
		All	Prompt	CR	SR	NR-Expert	NR-Crowdsourced
Hand-crafted	318	27	48	31	14	20	26
MI Conversations	4,365	31	31	33	27	NA	NA

Table 2. Dataset statistics for each data source. "NR" standards for non-reflection responses.

Complex Reflection (CR). Complex reflections are responses that add or infer something new from the client's statement. This may include naming a feeling or emotion that the client has not yet expressed, 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." shown in Table 1 is a complex reflection (i.e., high-quality response) as it brings attention to the client's traumatic experience, rather than merely rephrasing what was said. Complex reflections are considered a high-quality response.

Non-Reflection. These responses include unsolicited advice or questions asked when a reflection would have been a better response. NR are classified as low-quality and less desirable during the counselor learning process.

3.3 Non-expert Annotations.

To collect responses portraying beginner to nonexistent counseling skills, we obtain crowd-sourced annotations from lay individuals using Amazon Mechanical Turk. We believe that such responses provide realistic scenarios of what our system might encounter in counseling training. This step is inspired by our clinical collaborator's observation that providing unsolicited advice is a behavior frequently displayed when trainees are learning to craft reflections. This strategy allowed us to obtain diverse responses without the need for expert input. Additionally, we ensure response diversity by requesting three responses per prompt and annotations for unique workers. During the data collection, we showed a prompt to the worker and asked them to provide "advice" to the given scenario so their responses would likely contain directive rather than reflective language. Our annotation guidelines are shown below:

Task description. We are collecting responses to various conversational scenarios to help train a conversational AI system. Your task is to provide advice in response to a given situation or problem described.

Instructions. You will be presented with a description of a situation or problem that someone might be facing. Make sure you understand the context and the specific issue at hand. Write a response where you offer advice or suggestions on what the person should do. Think about what you would recommend if a friend came to you with this problem or situation, aiming to provide clear guidance.

We perform a validation step for crowd-sourced responses and reject them if they fail to follow the guidelines. Table 2 shows our final dataset statistics and the average number of tokens for each type of reflection in the dataset.

4 LANGUAGE MODELS FOR ASSESSING COUNSELOR REFLECTIONS

We explore two computational approaches for evaluating counselor reflections: (1) fine-tuned language models, and (2) larger, prompt-based language models with in-context learning. Our choice of these models is driven by their complementary strengths in processing and analyzing reflective listening within psychotherapy contexts. Fine-tuned models, e.g., RoBERTa, can be optimized in-house and thus offer practical advantages in terms of computational efficiency and applicability in training settings. Conversely, larger models like Flan [61], Mistral [23], and GPT-3.5 [7], leverage their extensive training on diverse datasets to provide a broader, more generalized understanding of counseling

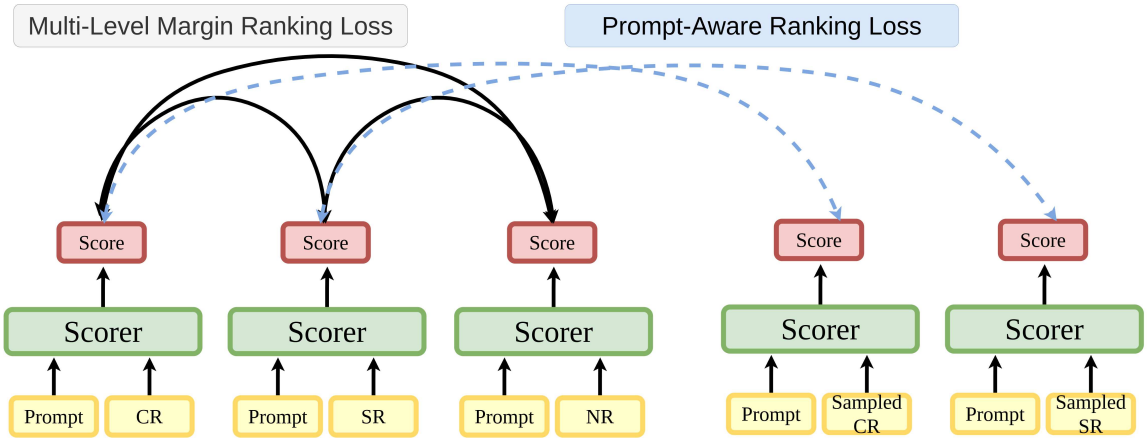


Fig. 2. Diagram of our model training framework. Our framework uses two types of contrasts between multiple levels of responses. The solid arrows represent the **multi-level margin ranking loss**, where different levels of responses to the same prompt are compared. This loss is designed so that the scorer learns to distinguish high-quality reflections from low-quality ones. The dashed arrows represent the **prompt-aware margin ranking objective**, where the comparison is on reflection pairs with different context prompts. This objective prevents the scorer from ignoring the context prompt when scoring.

language and empathy, essential for reflective listening. Their prompt-based interaction paradigm enables a flexible assessment of reflections, capturing a wide array of communicative subtleties. In this paper, we compare the efficacy and applicability of these strategies in evaluating counselor reflections to identify the most effective approach while providing scalable and effective feedback for counseling training.

4.1 Prompt-Aware margIn Ranking (PAIR)

Reflection scoring consists of assigning a score s between $[0, 1]$ to an interaction pair containing a client prompt p and a candidate reflection by a counselor r. While this task can be considered regression, obtaining ground truth labels for model training can be expensive and noisy, even with expert annotations. Instead, we develop a scoring framework inspired by contrastive and metric learning strategies. We pose the scoring problem as a learning-to-rank problem, in which the training data labels are pairwise relevance levels based on skill level, i.e., depending on whether the response is labeled as complex reflection (CR), simple reflection (SR), or non-reflection (NR). [9].

For our model backbone, we use Robustly Optimized BERT Pretraining (RoBERTa) [60], but note that our approach is flexible enough to be used with other transformer-based models. We use a cross-encoder that takes the concatenated sequence of a prompt and a response pair as input. Since this design choice allows us to model the interaction of prompt and response tokens directly, we classify our primary model as a cross-encoder-based model, following the characterization of the encoder provided by Humeau et al. [20]. We draw upon work from Lin et al. [30] to build our learning objectives, where binary contrastive estimations are computed between examples for consecutive reflection quality levels.

Multi-level Margin Ranking Objective. We designed a margin ranking loss term to ensure a distance gap between quality levels of reflections, taking inspiration from Lin et al. [30]. The ranking objective uses a margin parameter μ or 2μ, depending on the distance between examples being compared in the loss term.Hence, we use μ when the quality

gap is within one level, i.e., distinguishing between medium-quality and high-quality pairs (SR or 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. Similarly, $s(p, r_{CR}), s(p, r_{SR}), s(p, r_{NR})$ refer to the model predicted reflection score of the response r, given prompt 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 $\mathcal{L}_{gap}$, the model ignored the client prompt when making predictions, resulting in incorrect scoring for cases where responses are unrelated to the client prompt but follow a reflective language style. To address these cases, we designed a prompt-aware objective to penalize the model against such scenarios.

We thus simulate examples where the model receives a high or mid-quality response but is matched with a non-relevant prompt context and ensure that cases receive low scores. To provide these examples 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). The resulting pair should obtain a low score, even if the response itself is a valid reflection. 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 final model combines the two metric-learning-based objectives to enforce the correct ranking and prompt relevance. Figure 2 shows an overview of the training process. The scoring function is the transformer encoder model, followed by a pooling layer and a sigmoid activation where we combine the $\mathcal{L}_{gap}$ and $\mathcal{L}_{prompt}$ objectives with equal weights:

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

4.2 In-context Learning with Large Language Models (ICL-LLM)

In-context learning (ICL) with LLMs represents a paradigm shift in how AI systems are leveraged to understand and generate human-like text. This approach involves using prompts or examples within the input to guide the model's output and capitalize on the extensive pretraining of LLMs across vast and diverse datasets. Pretraining empowers LLMs with a broad understanding of language and knowledge and enables them to apply this generalized knowledge to specific tasks or queries without requiring task-specific fine-tuning. This is particularly advantageous as it circumvents the need for additional computational resources, domain-specific datasets, and the extensive time typically required for retraining models for new tasks.

ICL with LLMs can have an important impact on the counseling domain streamlining the process of training and assessing counseling skills by offering scalable and efficient means of providing feedback to practitioners.

We explore the use of ICL in evaluating the quality of reflections in counseling by analyzing the depth, empathy, and accuracy of the counselor's responses.

Reflection Classification with ICL. For each client prompt p and counselor response r pair (p, r), we seek to classify the response r into one of three categories: Complex Reflection (CR), Simple Reflection (SR), and Non-Reflection (NR). When crafting prompts for LLM-based reflection classification, it is important to concisely encapsulate the counseling context, provide clear definitions and examples for reflection categories (CR, SR, NR), and highlight the emotional and linguistic nuances of reflective listening. This ensures the model accurately classifies counselor responses by recognizing content and empathetic quality within a realistic therapeutic dialogue framework. To achieve this, we construct a system prompt by incorporating explicit task instructions, definitions of reflection types, and illustrative examples corresponding to each reflection quality class. The LLM prompt used during our experiments is shown in Appendix B.4.

5 EXPERIMENTS

5.1 Experimental Setup

PAIR model. Our models use the RoBERTa [32] architecture with pre-trained weights `mental-roberta-base` [22].

Our choice of pretrained weights is motivated by our domain being similar to 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 found that they improve overall performance.

Recent empirical findings also suggest that further fine-tuning on specialized domains improves performance on target tasks [17]. While this may come at the expense of performance degradation when applied to out-of-domain datasets, we prioritized performance over our domain (MI counseling) since we targeted a specific use case.

We implement our models using the PyTorch [44] and Huggingface Transformers [63] packages. For training, we use the Adam optimizer with a 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 efficiently fit the training data into our computing device, we subsample each data row into a smaller row. 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).

ICL-LLM Models. We experiment with three state-of-the-art LLMs, Flan-T5-XL, Mistral, and GPT-3. More specifically, we use Flan-t5-XL [61] a T5-based model fine-tuned for instruction-following tasks [48]; Mistral-7B-Instruct-v0.2 [23] a version of the Mistral-7B model that is optimized for adherence to specific instructions; and GPT-3.5-turbo-16k [7], a model with advanced linguistic capabilities. These models, designed for instruction-following tasks, offer nuanced, context-aware assessments in therapeutic settings [10], demonstrating a significant capacity for understanding and processing complex language tasks. Their application in the mental health domain showcases the potential of leveraging broad contextual awareness and specific guideline adherence for effective scoring in the health conversation domain [64].

However, black box LLMs such as GPT-3.5 are inherently limited by their closedness. Therefore, they require careful consideration before deployment in real settings, especially in education or counseling, where understanding

Metrics / Model	Naive Classifier*	Naive Classifier	Naive Regressor*	Naive Regressor	PAIR*	PAIR	Flan	Mistral	GPT-3.5
Recall@1	0.8952	0.8349	0.9174	0.5873	**0.9253**	0.6444	0.3325	0.0539	0.3786
Pearson	0.8713	0.7652	**0.8994**	0.7998	0.8722	0.7205	0.2914	0.1001	0.5330
Spearman	**0.8816**	0.7858	0.8784	0.7994	0.8811	0.7415	0.2766	0.0900	0.5324
Kendall's Tau	0.6955	0.5685	0.8653	0.7389	**0.8694**	0.7216	-0.0936	-0.4173	0.2243

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.

the decision process is important. Although researchers have devised different techniques such as chain-of-thought prompting [62] to improve reasoning and boost explainability, LLMs are still not good at offering explanations for their decisions in the clinical domain [42]. With this limitation in mind, we implement LLM-based prompting as a baseline, which requires no training data except for a few in-context examples.

Evaluation. To evaluate our models, we set aside 20% of our data as our test set. Our main performance metrics are recall@1, Pearson and Spearman, and Kendall's Tau correlation. 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 are arranged into ranked tuples according to the actual and predicted reflection levels, and we use the actual and predicted rankings to compute their correlation scores.

5.2 Baselines

We compare the different models against baselines sharing the same transformer encoder architecture and pre-trained weights as the PAIR model. We experiment with classifier and regressor models built using linear heads on top of the encoders. We use the same transformer model (`mental-roberta-base`) as our PAIR model and use the same parameter except for the prediction head. These baselines are primarily motivated by the need to separate the contribution of the PAIR loss from the contribution of the underlying model.

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

We also conduct evaluations using the same baselines trained on a prompt-aware loss term on top of original losses for a fair comparison. 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 for the different models and baselines on the set-aside test set while Table 4 shows the results when using the test set augmented with randomly-matched responses. In both tables, PAIR refers to our finetuned models trained with the complete set of our objectives, while * indicates that we remove the prompt-aware objective for ablation during training.

Metrics / Model	Naive Classifier*	Naive Classifier	Naive Regressor*	Naive Regressor	PAIR*	PAIR	Flan	Mistral	GPT-3.5
Recall@1	0.8952	0.8349	0.9174	0.5873	**0.9253**	0.6444	0.3325	0.0539	0.3786
Pearson	0.4892	0.6868	0.5317	0.6902	0.5108	**0.7396**	0.3644	0.0990	0.4609
Spearman	0.4896	**0.7227**	0.5018	0.6590	0.5001	0.6795	0.3394	0.0851	0.2766
Kendall's Tau	0.2397	0.4316	0.4539	0.5824	0.4485	**0.5940**	0.0371	-0.2757	0.1750

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.

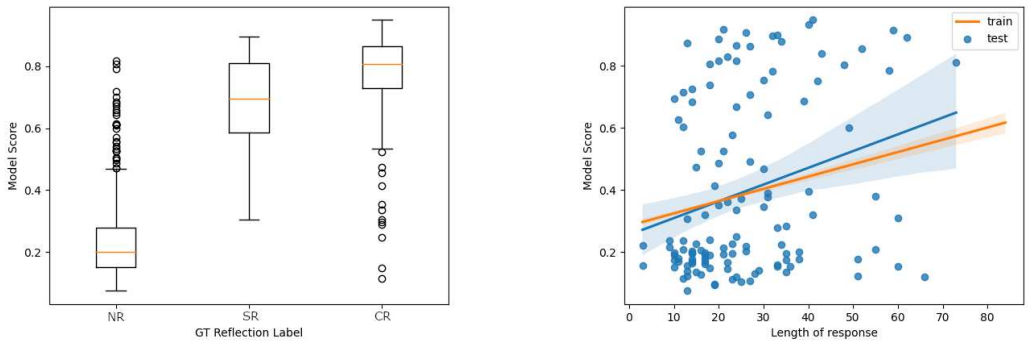
Prompt	Response	Model	Prediction
Have you ever tried to get off coke? Do you have any idea how hard it is to quit this stuff?	You believe that I may not understand exactly how hard your struggle is. (Complex Reflection)	PAIR	0.82
		Flan	0.5
		Mistral	0.0
		GPT-3.5	1.0
	Trying to get off coke has been really hard on you. (Simple Reflection)	PAIR	0.77
		Flan	0.5
		Mistral	0.0
		GPT-3.5	0.5
	You need to focus on sobriety and finding people that want to live clean and healthy. (Non-Reflection)	PAIR	0.11
		Flan	0.0
		Mistral	0.5
		GPT-3.5	0.0

Table 5. Sample Predictions from the models. The PAIR model outputs a continuous score, while other models' categorical labels (NR, SR, CR) are converted to scores for uniform presentation in this table.

In both sets of experiments, recall@1 results are identical, indicating that even after randomly matched responses are added, all the models can correctly identify responses with the highest reflection level (complex reflection). Moreover, we note that for the naive classifier models, Spearman correlation scores higher than Pearson correlation, likely because Spearman correlation measures monotonic relationships, and the naive classifier predictions contain frequent ties due to outputting a discrete categorization rather than continuous scores as the other models.

Comparison against baselines. The comparisons against baselines show exciting trends. First, when tested on data without randomly matched responses (Table 3), the best-performing baseline models (Naive Classifier, Naive Regressor) perform similarly to PAIR* (our PAIR model with the prompt-aware objective ablated). Although PAIR* outperforms baselines regarding recall@1 and Kendall's Tau, its Pearson and Spearman correlation coefficients are slightly worse than the naive models. However, results in Table 4 show that PAIR benefits from seeing mismatched responses. The experiments with the combined objectives show that PAIR outperforms the Naive Regressor model. When comparing the Naive Classifier and PAIR, we note that the Naive Classifier models are better for the recall@1 metric. However, we note that the two models are not directly comparable since they represent different frameworks of prediction and feedback. Additionally, because classifiers output a discrete label, they can be more robust against some noise in the output logits. Still, we argue that in scenarios where a continuous score is desired, our model is preferable to the classifier since it can provide more detailed feedback, better conveying the implicit preference ranking of different responses, as evidenced by its higher Kendall's Tau score.

PAIR ablations. We also evaluate our scorer models and baselines using ablations for the prompt-aware learning objective. In Table 3, when we measure the performance of our models on test cases where all prompt-response pairs are matched pairs (i.e., the response was in response to the matched prompt), prompt-aware models performed worse in



(a) Boxplot with Distribution of PAIR Predicted Scores

(b) Plot of Token Length and Model Predicted Scores

Fig. 3. Distribution of PAIR Predicted Scores and Relationship between Token Length and PAIR Predicted Scores.

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 suggests 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 shows the effectiveness of the prompt loss function in preventing the model from ignoring the prompt when scoring responses, hence making our model robust to cases where relevant responses are not guaranteed.

Comparison against LLMs. Finally, the exploration of LLMs, including Flan, Mistral, and GPT-3.5, revealed that while these models bring the advantage of extensive pre-training and a broad understanding of language, they trail behind the PAIR model. As shown in Table 3, LLMs—recall@1, Pearson, Spearman correlations, and Kendall's Tau—consistently underperform when compared to PAIR. These results highlight LLMs' challenges in specialized tasks without targeted fine-tuning, particularly in nuanced domains like psychotherapy, where understanding subtle language cues and context is crucial. Notably, domain-specific adaptations and the effectiveness of custom training objectives, such as the ones in PAIR, are important in achieving higher alignment with expert judgments of reflective listening.

6.1 PAIR Score Distribution Analysis

We visualize the distribution of predicted scores of prompt and response pairs in our test data using our main model, PAIR. Not surprisingly, Figure 3a shows it is harder for the model to distinguish between simple and complex reflections. This may be due to (1) the inherent difficulty in differentiating between simple and complex responses [45], and (2) the relatively fewer number of SR examples in the compiled dataset.

We observe that in our dataset, CRs have more tokens than SRs, and also CRs are longer than NRs on average as shown in Table 2. Given that we want our models to learn meaningful semantic and stylistic features of reflection rather than just relying on token response length, we plot the relationship between token length and reflection levels predicted by our final model.

Figure 3b shows the scatterplot of response length in # of tokens and predicted reflection score, with two regression lines for the training set (blue) and testing set (orange) results. The regression line for the training set is computed using

model scores, while for the testing set, ground truth judgments are converted to a continuous score identical to that used to train the Naive Regression model. The large dispersion in the distribution indicates that the response length is only slightly positively correlated with the PAIR scorer, ensuring that the model is not making spurious generalizations from the response length.s

7 DEPLOYMENT OF PAIR TO PRODUCE FEEDBACK IN COUNSELING TRAINING

In addition to PAIR on annotated data, we also deploy it in a real-life education setting – a graduate-level MI training course taught by one of the authors of this paper [2]. The graduate students in this class are training to be MI counselors, and our objective in deployment is to verify our system's usefulness in providing feedback to trainee counselors. Several factors drove the choice of PAIR over larger language models (LLMs). Firstly, during our experiments, PAIR outperformed other approaches and showed (Section 6), a better alignment with an expert evaluation of reflective listening. Secondly, since PAIR is a more contained model, it reduces the risk of data exposure compared to LLM use. Furthermore, Regarding latency and reliability, a locally-run, smaller PAIR model offers lower response time and more predictable and controllable output generation.These advantages make PAIR a better choice for embedding within larger systems to enhance counseling skills training, aligning with frameworks suggested in prior research [51] while also ensuring a practical, secure, and effective learning tool for MI training.

7.1 Model Robustness for Deployment Safety

We carried out comprehensive robustness evaluations to ensure the safety and reliability of the PAIR model, especially in handling a wide array of student inputs. This precautionary measure was conducted to confirm that the model could effectively manage the diverse and complex nature of student responses in educational settings without compromising its integrity or the quality of its output.

Beyond measuring the model performance on the test data, we also evaluate the robustness of our models using the Checklist framework by Ribeiro et al. [50]. We focus our tests on the counselor's response rather than the client's prompt since the main goal is to provide feedback on trainee responses. We assess whether PAIR leverage both stylistic cues or features of counselor language and semantics. We conducted three main types of tests, as described below.

Minimum Functionality Test (MFT).. Evaluates whether the model correctly scores longer utterances that do not contain reflection language that is also unseen during training time. We also check whether phrases or expressions frequently used in reflections e.g., reflection starters such as: "it sounds like ..." are correctly identified as non-reflective when used in isolation.

Directional Expectation Test (DIR).. We test whether paraphrased versions of counselor language receive similar scores as the originals. For this task, we used an off-the-shelf paraphrase model.[3]

Invariance Test (INV).. We apply noise to the counselor's language (typos, punctuation, contractions) and measure the change in reflection scores. We also test if the resulting score remains low when reflective phrases, e.g., reflection starters, are inserted into non-reflective counselor language.

For each test, we measure the amount of change in score to indicate scoring failure. For MFT tests, we use dialog turns extracted from the MITI sessions. Results are shown in Table 6.

[2] link
[3] https://huggingface.co/tuner007/pegasus_paraphrase

Test Type	Description	#Samples	Error Rate
MFT_LONG	Lengthy NR responses receive low scores.	4,225	0.02
MFT_FRAME	Isolated reflective phrases receive low scores.	36	0.07
DIR_PARAPHRASE	Paraphrases receive similar scores to the original reflection.	495	0.11
INV_NOISE	Typos, contractions, and punctuation errors do not affect the score.	422	0.16
INV_FRAME	Reflection phases followed by low quality (NR) response receive low scores.	315	0.03

Table 6. Failure rate for PAIR error checklist

The results show a low error rate in all the conducted tests, suggesting that PAIR is good at distinguishing a reflection-sounding language from semantically genuine reflections (MFT_LONG, MFT_FRAME, and INV_FRAME) and further showing robustness to different types of input perturbation.

7.2 Deployment Setup

To evaluate our model in a real-world setting, we collaborated with psychotherapy faculty at the University of Michigan School of Public Health. We deployed PAIR in a graduate-level MI training course[4]. The system was used by students as part of their course assignments.

System Design and Implementation. We deploy PAIR in a web-based application to provide real-time scoring feedback to students while learning to create reflective responses to a given client prompt. The system is implemented as a web server using Nginx,[5] Gunicorn,[6] and the Flask[7] web framework and is run on a secure server. Running PAIR for $\tilde{3}0$ prompt evaluation and response pairs and providing feedback takes less than one second. The system presents five client prompts at a time, but students only need to provide at least one response to receive feedback. Students are allowed to complete their assignment at their own pace as the system is able to save and retrieve their work at any time. After the assignment is submitted, the model is run in the server and students are presented with detailed feedback to each response, including two ground truth high-quality reflections for each prompt, and the model predicted scores.

Participants. Our participant pool consisted of 30 students enrolled in a graduate-level psychotherapy class. The students used our system to complete three assignments that required them to practice their reflective skills. Over 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. A screenshot of our web interface can be found in Figure 5.

7.3 Usability Evaluation

We designed a 5-point Likert survey to assess the perceived accuracy and usability of our system that was presented to students after they submitted their assignments. Figure 4 shows the survey questions covering model error and system usability and the distribution of student responses. Overall, our system received a positive assessment inaccessibility and performance aspects. We also asked users to submit free-form text feedback to learn more about their experiences while using the system. Among the submitted comments, positive answers focus on how the application allowed them

[4] https://sph.umich.edu/academics/courses/syllabi/HBEHED671.pdf
[5] https://www.nginx.com
[6] https://gunicorn.org
[7] https://flask.palletsprojects.com/en/2.1.x/

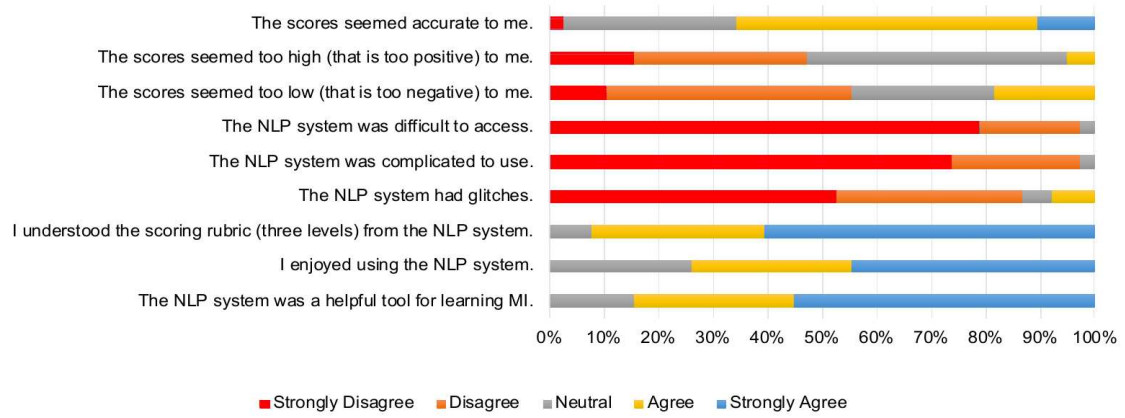


Fig. 4. User survey results on a 5-point Likert scale. For comparing answers in a unified positive scale, questions 5-9 were negated.

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 7. Confusion matrix of the model predictions on student submissions.

to practice more and build confidence. At the same time, negative feedback is usually concerned with functionality aspects such as saving and loading their work.

7.4 Evaluation on User-generated Responses

We evaluate PAIR performance in scoring actual responses from students using expert annotations. One of the course instructors reviewed students' responses and annotated them as CR, SR, or NR. We use these annotations as our ground truth set to evaluate the performance of our model. To compare scores and categorical labels, we first convert the PAIR-predicted score into a discrete label using the following mapping: $\{\text{CR} : [0.7, 1.0], \text{SR} : [0.3, 0.7], \text{NR} : [0.0, 0.3]\}$. We then evaluate the system performance regarding accuracy and correlation with the ground truth. Results are shown in Table 7.

As indicated by the confusion matrix and accuracies in Table 7, our model performs the 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 for this class, and hence the low false positive rate shown by the system is aligned with this design objective.

7.5 Quantitative Error Analysis

To further understand the system performance, we conduct an error analysis based on the false positive and negative rate. False positives occur when the model provides high scores for responses that have low to nonexistent reflective language. Conversely, false negatives happen when the model assigns low scores to responses with highly reflective language. According to MI experts feedback, false negatives are considered more detrimental for MI training as they

may inadvertently reinforce inadequate counseling responses, potentially hindering the development of effective reflective listening skills. Table 8 shows a false positive and a false negative example. In the first example, the model overestimates the reflection score. The response empathizes with the client's frustration but fails to explore deeper feelings or alternative solutions, which would be characteristic of a higher-quality reflection. In the second example, the response shows a nuanced understanding of the clients' situations and express empathy, yet it was scored lower than their qualitative content suggests. These discrepancies highlight the challenges in accurately assessing reflections.

Prompt	Response	Score
False positive		
Well, how am I supposed to cook red beans and rice if I can't use sausage because of salt?	You are angry because you feel like something you enjoy has been taken away from you. You aren't sure yet how to handle that loss of freedom.	0.41
False negative		
I tried giving my kids fruit for snack, if they don't have their cookies, they make a huge fuss. They expect sweets after school, and I can't stand the sound of their whining when they don't get what they want. Plus, I kind of like baking homemade treats.	You want to make beneficial dietary changes for your children, but their poor behavior makes it difficult for you to follow through. You enjoy preparing food for them and wish you could find something they would not complain about that has less sugar.	0.21

Table 8. Sample false positive and negative examples

8 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 on 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 in future work.

9 LIMITATIONS

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

First, our PAIR model is finetuned using data manually 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 our data and models trained using counseling data collected in the wild.

Second, in this study, we confined our exploration of large language models (LLMs) to in-context learning using general-domain frameworks, opting not to finetune LLMs on our specific dataset [64]. This approach is motivated by prior research indicating that employing LLMs through in-context learning, without fine-tuning, yields superior outcomes for therapist behavior classification [10]. Additionally, our research focuses on comparing two distinct paradigms: one utilizing a smaller model with a highly specialized PAIR loss, and the other employing expansive,

generalized LLMs, which are increasingly recognized for their versatility across a broad spectrum of NLP tasks. Moving forward, we intend to study the potential of LLMs to be effectively fintuned and trained on our dataset, aspiring to combine the strengths of both targeted and generalized approaches.

Finally, the PAIR 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 plan 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' responses.

10 CONCLUSION AND FUTURE WORK

In this work, we explored the application of computational models, from RoBERTa variants fintuned on custom losses to sophisticated large-scale language models like Flan, Mistral, and GPT-3.5, for the assessment of reflective listening in mental health counseling. Through the use of the PAIR dataset, we have demonstrated the effectiveness of a fintuned RoBERTa model equipped with a Prompt-Aware margIn Ranking (PAIR) learning objective in providing targeted feedback for the development of counseling skills. Our findings reveal that the comparatively smaller, fintuned PAIR model surpasses the performance of more generalized LLMs across an array of evaluation metrics. The PAIR framework 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.

We make our data available at https://lit.eecs.umich.edu/downloads.html#PAIR.

ACKNOWLEDGEMENTS

The authors would like to thank 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 a National Science Foundation award (#2306372). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

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

[2] Victor Ardulov, Torrey A. Creed, David C. Atkins, and Shrikanth Narayanan. 2022. Local dynamic mode of Cognitive Behavioral Therapy. https://doi.org/10.48550/ARXIV.2205.09752

[3] Miller W. R. Arkowitz, H. and S Rollnick. 2015. Motivational interviewing in the treatment of psychological problems, 2nd edition.

[4] 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 (2007), 193–9.

[5] Alain Braillon and Françoise Taiebi. 2020. Practicing "Reflective listening" is a mandatory prerequisite for empathy. *Patient Education and Counseling* 103, 9 (2020), 1866–1867. https://doi.org/10.1016/j.pec.2020.03.024

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

[8] Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5599–5611. https://doi.org/10.18653/v1/P19-1563

[9] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to Rank: From Pairwise Approach to Listwise Approach. *Proceedings of the 24th International Conference on Machine Learning* 227, 129–136. https://doi.org/10.1145/1273496.1273513

[10] Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A Computational Framework for Behavioral Assessment of LLM Therapists. arXiv:2401.00820 [cs.CL]

[11] Munmun De Choudhury, Sachin R. Pendse, and Neha Kumar. 2023. Benefits and Harms of Large Language Models in Digital Mental Health. arXiv:2311.14693 [cs.CL]

[12] Neo Christopher Chung, George Dyer, and Lennart Brocki. 2023. Challenges of Large Language Models for Mental Health Counseling. arXiv:2311.13857 [cs.CL]

[13] Debadutta Dash, Rahul Thapa, Juan M. Banda, Akshay Swaminathan, Morgan Cheatham, Mehr Kashyap, Nikesh Kotecha, Jonathan H. Chen, Saurabh Gombar, Lance Downing, Rachel Pedreira, Ethan Goh, Angel Arnaout, Garret Kenn Morris, Honor Magon, Matthew P Lungren, Eric Horvitz, and Nigam H. Shah. 2023. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. arXiv:2304.13714 [cs.AI]

[14] 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 (2021). arXiv:2102.11573 https://arxiv.org/abs/2102.11573

[15] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[16] Jacques Gaume, Gerhard Gmel, Mohamed Faouzi, and Jean-Bernard Daeppen. 2009. Counselor skill influences outcomes of brief motivational interventions. *Journal of Substance Abuse Treatment* 37, 2 (2009), 151–159. https://doi.org/10.1016/j.jsat.2008.12.001

[17] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

[18] Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. 2022. Structured Prompting: Scaling In-Context Learning to 1, 000 Examples. *ArXiv* abs/2212.06713 (2022). https://api.semanticscholar.org/CorpusID:254591686

[19] 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:2305.08982 [cs.HC]

[20] 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*.

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

[22] 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*.

[23] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL]

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. *ArXiv* abs/2004.11362 (2020). https://api.semanticscholar.org/CorpusID:216080787

[25] Brian Kulis. 2013. Metric Learning: A Survey. *Found. Trends Mach. Learn.* 5 (2013), 287–364. https://api.semanticscholar.org/CorpusID:55485900

[26] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2022. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In *Conference of the European Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:252907607

[27] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. *ArXiv* abs/2307.11991 (2023). https://api.semanticscholar.org/CorpusID:260125719

[28] Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer. In *North American Chapter of the Association for Computational Linguistics*.

[29] Inna Lin, Lucille Njoo, Anjalie Field, Ashish Sharma, Katharina Reinecke, Tim Althoff, and Yulia Tsvetkov. 2022. Gendered Mental Health Stigma in Masked Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2152–2170. https://doi.org/10.18653/v1/2022.emnlp-main.139

[30] 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*.

[31] Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. 2021. DialogueCSE: Dialogue-based Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2396–2406. https://doi.org/10.18653/v1/2021.emnlp-main.185

[32] 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 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[33] Sarah Lord, Elisa Sheng, Zac Imel, John Baer, and David Atkins. 2014. More Than Reflections: Empathy in Motivational Interviewing Includes Language Style Synchrony Between Therapist and Client. *Behavior Therapy* 46 (11 2014). https://doi.org/10.1016/j.beth.2014.11.002

[34] Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are Emergent Abilities in Large Language Models just In-Context Learning? *ArXiv* abs/2309.01809 (2023). https://api.semanticscholar.org/CorpusID:261531236

[35] Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness Transfer: A Tag and Generate Approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 1869–1881. https://doi.org/10.18653/v1/2020.acl-main.169

[36] William R Miller and Stephen Rollnick. 2013. *Motivational interviewing: Helping people change, Third edition.* The Guilford Press.

[37] 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10289–10302. https://doi.org/10.18653/v1/2023.findings-emnlp.690

[38] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *ArXiv* abs/2202.12837 (2022). https://api.semanticscholar.org/CorpusID:247155069

[39] Adam S. Miner, Nigam Shah, Kim D. Bullock, Bruce A. Arnow, Jeremy Bailenson, and Jeff Hancock. 2019. Key Considerations for Incorporating Conversational AI in Psychotherapy. *Frontiers in Psychiatry* 10 (2019). https://doi.org/10.3389/fpsyt.2019.00746

[40] 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 (01 2016). https://doi.org/10.1016/j.jsat.2016.01.001

[41] 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 (01 2016). https://doi.org/10.1016/j.jsat.2016.01.001

[42] Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. arXiv:2401.06775 [cs.CL] https://arxiv.org/abs/2401.06775

[43] Jihyun Park, Abhishek Jindal, Patty B. Kuo, Michael J Tanana, Jennifer Elston Lafata, Ming Tai-Seale, David C. Atkins, Zac E. Imel, and Padhraic Smyth. 2021. Automated rating of patient and physician emotion in primary care visits. *Patient education and counseling* (2021).

[44] 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 *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[45] 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*. Association for Computational Linguistics, San Diego, CA, USA, 42–51. https://doi.org/10.18653/v1/W16-0305

[46] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. 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)*. Association for Computational Linguistics, Vancouver, Canada, 1426–1435. https://doi.org/10.18653/v1/P17-1131

[47] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J. Goggin, and Delwyn Catley. 2017. 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*. Association for Computational Linguistics, Valencia, Spain, 1128–1137. https://aclanthology.org/E17-1106

[48] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]

[49] 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 (02 2007), 191–209. https://doi.org/10.1176/appi.psychotherapy.2007.61.2.191

[50] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4902–4912. https://doi.org/10.18653/v1/2020.acl-main.442

[51] 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* (2021).

[52] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2022. Human-AI Collaboration Enables More Empathic Conversations in Text-based Peer-to-Peer Mental Health Support. arXiv:2203.15144 [cs.CL]

[53] Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *EMNLP*.

[54] Ashish Sharma, Kevin Rushton, Inna Lin, David Wadden, Khendra Lucas, Adam Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 9977–10000. https://doi.org/10.18653/v1/2023.acl-long.555

[55] 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)*. Association for Computational Linguistics, Dublin, Ireland, 3096–3107. https://doi.org/10.18653/v1/2022.acl-long.221

[56] 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*. Association for Computational Linguistics, 1st virtual meeting, 10–20. https://aclanthology.org/2020.sigdial-1.2

[57] K. Singhal, Shekoofeh Azizi, Tao Tu, Said Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather J. Cole-Lewis, Stephen J. Pfohl, P A Payne, Martin G. Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, P. A. Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Greg S. Corrado, Yossi Matias, Katherine Hui-Ling Chou, Juraj Gottweis, Nenad Tomaev, Yun Liu, Alvin Rajkomar, Joëlle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *Nature* 620 (2022), 172 – 180. https://api.semanticscholar.org/CorpusID:255124952

[58] Elizabeth Stade, Shannon Stirman, Lyle Ungar, H. Schwartz, David Yaden, João Sedoc, Robert DeRubeis, Robb Willer, and johannes Eichstaedt. 2023. Artificial intelligence will change the future of psychotherapy: A proposal for responsible, psychologist-led development. https://doi.org/10.31234/osf.io/cuzvr

[59] 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 (July 2019), e12529. https://doi.org/10.2196/12529

[60] 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*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[61] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. [n. d.]. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

[62] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] https://arxiv.org/abs/2201.11903

[63] 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*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[64] Xuhai Xu, Bingsheng Yao, Yu Dong, Hongfeng Yu, James A. Hendler, Anind K. Dey, and Dakuo Wang. 2023. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. https://api.semanticscholar.org/CorpusID:260203216

[65] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. On the Evaluations of ChatGPT and Emotion-enhanced Prompting for Mental Health Analysis. *ArXiv* abs/2304.03347 (2023). https://api.semanticscholar.org/CorpusID:258040984

[66] Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020. Quantifying the Causal Effects of Conversational Tendencies. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 131 (oct 2020), 24 pages. https://doi.org/10.1145/3415202

[67] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking Large Language Models for News Summarization. arXiv:2301.13848 [cs.CL]

A REFLECTION LEVELS AND EMPATHY COMMUNICATION

One of the core tenets of MI counseling is expressing empathy [36, 49]. However, the notion of desirable empathetic behavior is more specified than general empathetic behavior, since MI emphasizes accurately reflecting client sentiments and concerns, rather than expressing similar feelings or offering support [5, 33, 41]. Sharma et al. [53] studies different communication mechanisms for empathetic communication in the text-based mental support domain. In this experiment, we explore the relationship between empathetic communication and counselor reflection in MI.

To study the relationship between reflective verbal behavior and empathy in mental health exchanges, we adopt Sharma et al. [53]'s computational framework for analyzing expressed empathy. Their EPITOME framework studies three communication mechanisms of empathy: explorations, interpretations, and emotional reactions, each of which focuses on expressing emotions, conveying understanding or sympathy, and seeking further information, respectively.

Reflection Source / Empathy Mechanism	Emotional Reactions		Interpretations		Explorations	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Ground Truth	-0.1454	-0.1533	0.0156*	0.0192*	-0.3308	-0.3426
Model Score	-0.1355	-0.1319	0.0329*	0.0372*	-0.3639	-0.3476

Table 9. Correlation between reflection levels (both ground truth labels and model predicted scores) and EPITOME-detected communication mechanisms over the entire expert annotated dataset. * indicates results with p-value > 0.05.

We show the correlation between reflection levels and epitome communication mechanisms by computing the Pearson and Spearman correlations between the grown truth and model reflection levels, and the empathy communication levels predicted from the EPITOME model proposed in [53]. We train the model on the Reddit mental health dataset used in the paper to derive the communication mechanism scores and present the results in Table 9. We note a domain shift of the dataset used in training (Reddit posts and replies) and testing (client-counselor exchanges), which likely leads to the suboptimal performance of the trained model on our dataset. However, we observe that the EPITOME model, despite the domain shift, still provides meaningful insights into the relationship between reflection levels and empathy communication mechanisms.

We observe that emotional reactions and explorations are negatively correlated with reflection levels, indicating that the reflected verbal behavior of MI counselors is distinct from general empathetic behavior in mental health exchange. For example, explorative counselor utterances are often in question form, which is low in reflection. Moreover, although the correlation results are not statistically significant, the small but positive relationship between reflection and interpretation coincides with the characterization that reflection is a statement that expresses understanding. Finally, we note that this result is limited by model error and domain mismatch between the Reddit dataset used in EPITOME and our MI dataset.

B ADDITIONAL INFORMATION ON OUR USER STUDY

B.1 Obtaining Consent from Participants

Before they could access the assignments, participants were asked to read and sign an informed consent form, which informed them that their submissions would be securely stored and 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.

B.2 Data Anonymization and Protection

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

B.3 Annotator Guideline for Crowdsourced Data

```
Project Overview
We are collecting responses to various scenarios to help train a
    conversational AI. Your task is to provide advice in response to a given
    situation or problem described in a client prompt.

Read the Client Prompt Carefully: You will be presented with a description of
    a situation or problem that someone might be facing. Make sure you
    understand the context and the specific issue at hand.

Provide Your Advice: Based on the prompt, write a response where you offer
    advice or suggestions on what the person should do. Think about what you
    would recommend if a friend came to you with this problem, aiming to
    provide clear, directive guidance.
```

B.4 LLM Prompt

```
llm_prompt = f"""Your task is to score a counselor response to the client's
    prompt \
according to the categories of Complex Reflection, Simple Reflection, and
    Non-Reflection in motivational interviewing.

Complex Reflection goes deeper than what the client has directly expressed,
    offering a new perspective or insight. \
It often involves paraphrasing or expanding on the client's feelings or
    thoughts in a way that suggests a deeper understanding.
Simple Reflection involves mirroring or paraphrasing the client's statement
    without adding significant new meaning or interpretation.
Non-Reflection responses do not mirror or expand upon the client's \
feelings or statements but may offer advice, provide information, ask a
    question, or change the subject.

1. Complex Reflection Example:
Prompt: {complex_prompt}
Response: {complex_response}
```

```
2. Simple Reflection Example:
Prompt: {simple_prompt}
Response: {simple_response}

3. Non-Reflection Example:
Prompt: {non_prompt}
Response: {non_response}

Target Client Prompt: {prompt}
Target Counselor Response: {response}
Output 1.0 for a response that is a Complex Reflection, 0.5 for a Simple
    Reflection, and 0.0 for a Non-Reflection.
Score:
"""
```

B.5 Web Interface

Our web interface used for the user study is shown in Figure 5.

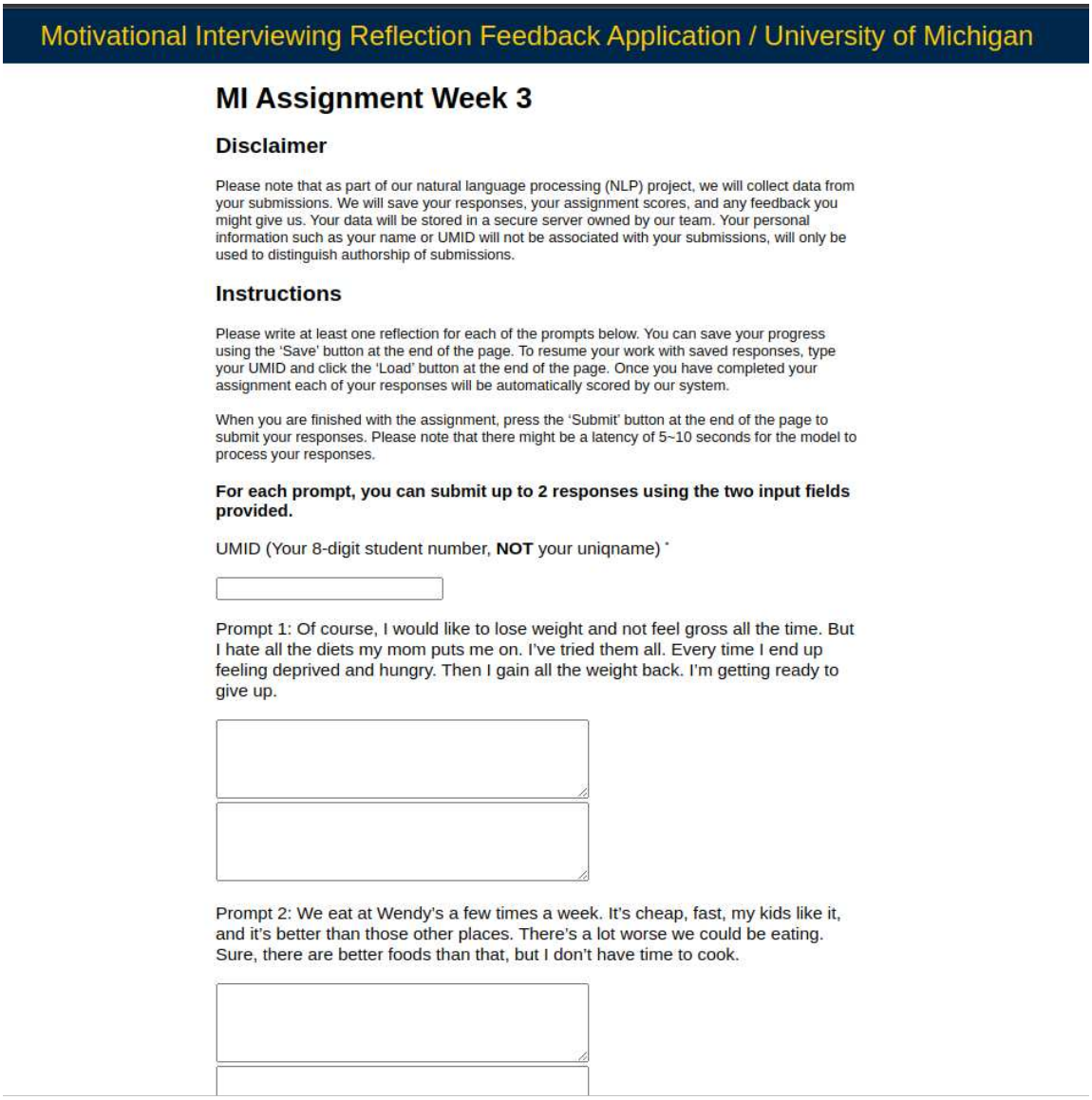


Fig. 5. A view of our web interface