Motivational Interviewing Transcripts Annotated with Global Scores

Ben Cohen, Moreah Zisquit, Stav Yosef, Anat Klomek Brunstein, Doron Friedman and Kfir Bar
Reichman University, Israel
{ben.cohen, stav.yosef, moreah.zisquit}@post.runi.ac.il
{bkanat, doronf, kfir.bar}@runi.ac.il

Abstract

Motivational interviewing (MI) is a counseling approach that aims to increase intrinsic motivation and commitment to change. Despite its effectiveness in various disorders such as addiction, weight loss, and smoking cessation, publicly available annotated MI datasets are scarce, limiting the development and evaluation of MI language generation models. We present MI-TAGS, a new annotated dataset of MI therapy sessions written in English collected from video recordings available on public sources. The dataset includes 242 MI demonstration transcripts annotated with the MI Treatment Integrity (MITI) 4.2 therapist behavioral codes and global scores, and Client Language EAsy Rating (CLEAR) 1.0 tags for client speech. In this paper we describe the process of data collection, transcription, and annotation, and provide an analysis of the new dataset. Additionally, we explore the potential use of the dataset for training language models to perform several MITI classification tasks; our results suggest that models may be able to automatically provide utterance-level annotation as well as global scores, with performance comparable to human annotators.

Keywords: Motivational Interviewing, computational mental health, MITI, CLEAR

1. Introduction

There is a recent increase in behavioral health concerns, including substance abuse, smoking cessation, eating disorders, and more (Maraz et al., 2021). Due to the common perception that behavior can be changed, behavioral counseling has emerged as an important tool for identifying unhealthy or self-destructive behaviors and possibly changing them (Rollnick et al., 2008). Motivational Interviewing (MI) is a psycho-therapeutic technique designed to aid individuals in addressing their ambivalence toward behavior change, employing a collaborative and client-centered approach (Miller and Rollnick, 1993).

Despite much research that points to its efficacy, it remains almost impossible to provide regular and immediate performance evaluations for therapists in their clinical practice. Behavioral coding—the process of observing therapist behaviors and assessing their skills by listening to audio recordings and/or reading session transcripts (Bakeman and Haynes, 2015)—is both time-consuming and cost-prohibitive in real-world settings.

There are currently two main coding systems for MI. Motivational Interviewing Skills Code (MISC)[1] which is commonly used to capture both therapist and client behaviors, and Motivational Interviewing Treatment Integrity (MITI),[2] which captures only the therapist behavior. There is an additional coding system that only captures client behavior, called Client Language Easy Rating (CLEAR).[3] It has been reported (Moyers et al., 2005) that after intensive training and supervision that lasts on average three months, a proficient coder would require up to two hours to code just one 20-minute session. Figure 1 showcases a snippet from an MI session.

T: [Persuade] `Ok, so in general we just want to work toward that.`
C: [Neutral] `Uh huh.`
T: [Giving Information] `So the most important component in consistent exercise is, "Does it fit into my routine". And as you know we go through lifestyle changes and you know in my case from having a family to not having a family... So do you want to brainstorm ways to keep it in, once you've finished your traveling... Let's figure out how you can do it. So that it really fits in logistically.`
C: [Change Talk] `If I carve out a specific time of day you know, and try and get my husband in on it so I have a partner... that would be the best. You know that's how it worked for me before.`

Figure 1: An excerpt from a session transcript, annotated with MITI codes for the therapist's utterances and CLEAR codes for the client's responses.

[1] https://casaa.unm.edu/tools/misc.html
[2] https://motivationalinterviewing.org/sites/default/files/miti4_2.pdf
[3] https://casaa.unm.edu/assets/docs/clear.pdf

LREC-COLING 2024, pages 11642–11657
20-25 May, 2024.

Natural-language-processing (NLP) technology has the potential to provide a solution to this coding challenge, yet the application of NLP to therapy, and more specifically to MI, has been impeded by the absence of publicly accessible MI sessions. Several published papers present MI datasets (Wade et al., 2009; Pérez-Rosas et al., 2016, 2017; Wu et al., 2022). However, most data in this domain are not shared due to ethical and privacy concerns.

Our work makes two main contributions. First, we introduce an annotated dataset comprised of 242 spoken dialogue MI sessions sourced from public domains. These primarily include demonstration videos by professional counselors and MI role-play counseling sessions by psychology students. Each video portrays different speakers, and the conversations cover various health topics such as smoking cessation, alcohol consumption, substance abuse, weight management, and medication adherence. We carefully annotated the data for MITI, including global scores, as well as for CLEAR. Since most of the existing data sets are not available due to privacy issues, we hope that this new dataset will become a benchmark for the scientific community to build and test new automatic MI coders or automatic behavioral health caregivers. Additionally, we provide an automated MITI feedback system for evaluating counselor performance, which can be used to code and evaluate a conversation or a fragment of a conversation. Hence, the current work will promote data availability and explore the automation of coding MI (Miller and Rollnick, 1993).

1.1. Behavioral Coding for Motivational Interviewing

Behavioral coding systems or annotation schemes provide comprehensive information about the process of motivational interviewing, and can provide feedback to counselors learning MI. In selecting an annotation scheme for analyzing MI sessions, we opted for MITI due to its popularity and simplicity. However, the MITI codes are assigned only to the therapist's utterances. Therefore, in order to code the client responses, we incorporated the CLEAR system as well.

1.1.1. Motivational Interviewing Treatment Integrity

The Motivational Interviewing Treatment Integrity (MITI) is a behavioral coding system that provides an answer to how well or poorly a practitioner is using MI. MITI also yields feedback that can be used to increase clinical skills in the practice of motivational interviewing. MITI is intended to be used: i) as a treatment integrity measure for clinical trials of motivational interviewing, and ii) as a means of providing structured, formal feedback about ways to improve practice in non-research settings.

The MITI coding system includes four global scores, each assigned once for the entire session using a 5-point Likert scale. Additionally, there are ten counselor behavioral codes assigned to each utterance made by the therapist. It is recommended for use with a 20-minute session segment, but it can also be applied to shorter or longer segments. If the segments are shorter or longer than 20 minutes the manual recommends interpreting the global scores with caution. The complete guidelines of MITI can be found here: `https://motivationalinterviewing.org/sites/default/files/miti4_2.pdf`.

1.1.2. Client Language Easy Rating (CLEAR)

The CLEAR system is designed to classify client responses into either change talk (CT) or counterchange talk (CCT). It emphasizes the types of in-session client language that have historically predicted future change (or lack thereof). The advantages of CLEAR include its simplicity, ease of training and use, and its capacity to compute the percentage-change-talk variable. The full guidelines can be found here: `https://casaa.unm.edu/assets/docs/clear.pdf`.

2. Related Work

There are currently only three publicly accessible MI corpora that have been released. Pérez-Rosas et al. (2019) gathered MI videos from publicly available platforms like YouTube and Vimeo and released a dataset in which each session is annotated as either a high-quality or low-quality counseling session. The transcripts were generated using a speech recognition system, which may have some minor inaccuracies. Currently, the annotated version of the dataset is not publicly available.

Welivita and Pu (2022) gathered data by scraping conversational data from peer support platforms such as Reddit and CounselChat, resulting in MITI labeled data of very short written language fragments with a few exchanges, that is, similar to an online forum discussion rather than a real therapy session.

More recently, Wu et al. (2022) presented a notable improvement over the previous datasets in two critical aspects. The most important aspect is that the dataset is now publicly accessible, making it widely available for research and analysis. Additionally, the transcription and annotation processes were conducted by professionals with meticulous attention to detail, ensuring a high level of accuracy and reliability. The sessions they annotated were sourced from publicly available resources.

Cao et al. (2019) tackled the task of real-time behavioral code classification. Their model predicts the next response's code based on session history, guiding therapists towards appropriate MI interventions. Using the MISC annotation scheme, they compiled annotated sessions from various sources. They employed simple recurrent neural networks with a word-level attention mechanism. Their findings indicate the model's potential to assist therapists in determining if an empathic response is needed next, during treatment.

Another type of studies try to evaluate specific targets in MI, for example: predicting empathy (Pérez-Rosas et al., 2017), analysing behavior change talk (Tavabi et al., 2021), and classifying high/low quality counseling conversations Pérez-Rosas et al. (2019). More recent studies have created a fully automatic evaluation system for MI (Flemotomos et al., 2021; Imel et al., 2019) and (Hershberger et al., 2021) which aims to help trainers with immediate feedback on the trainee skills.

Building on the promising use of annotated MI sessions in previous studies for predicting MITI or MISC labels, we introduce a new dataset comprising fully transcribed and annotated MI sessions based on MITI and CLEAR. Our collection includes 242 sessions; 102 are entirely new, while the remainder are corrected and annotated versions of sessions that were also included in previous datasets. For the first time, we manually assign the four MITI global scores to nearly all sessions in our dataset. This pioneering effort paves the way for developing automatic MITI evaluators for training sessions, potentially enhancing real-time practice.

We opted for MITI and CLEAR over MISC primarily due to the complexity of the MISC scheme. MITI is more straightforward, needing only a single read-through and annotation of each session, whereas MISC requires three. Pairing MITI with CLEAR allows us to assess the behaviors of both the therapist and the client, much like MISC does.

3. Method

3.1. Data Collection

We began our data collection on January 2022 with the transcribed video sessions published by Pérez-Rosas et al. (2019), sourced from online public platforms which at that time was the only dataset available to us. We observed that some transcripts are duplicates, others contain errors—likely resulting from the use of automatic transcript software—and some do not match their corresponding videos. We removed duplicates and sessions that were not specifically MI, as determined by one of the co-authors, who is a clinical psychologist. We corrected all transcripts by listening to the original videos. Sessions for which we could not locate the videos were removed. As previously mentioned (Pérez-Rosas et al., 2016), some videos contain the words 'good' or 'bad' in the title (for example, 'Mr. Wilson, Part 1 (not-so-good example)'), but we do not include this differentiation in our dataset. Including both "good" and "bad" session transcripts is encouraged for creating a comprehensive dataset, as it is more likely to capture the full range of therapist and client behavioral codes, as well as a complete range of global scores.

From the 259 original sessions provided by Pérez-Rosas et al. (2019), we retained only 140 that had been fully validated and corrected by us. The remaining sessions could not be validated due to unavailable videos or severe discrepancies between the transcript and the original video. To expand our dataset, we searched platforms such as YouTube, Vimeo, as well as the professional archive Alexander Street,[4] using the query "motivational interviewing". Eight additional sessions were taken from The Center on Alcohol, Substance Use And Addictions (CASAA),[5] of The University of New Mexico. All together this search yielded 102 new MI sessions, bringing our total to 242 MI session transcripts. The newly added sessions primarily feature exemplary MI counseling practices, as they are designed to serve as tutorials for those aspiring to improve their MI skills.

3.2. Data Verification and Pre-processing

For the 102 new videos, we generated transcripts as follows: For YouTube video, we downloaded the YouTube generated captions, and for Vimeo videos, we extracted the audio signal from the video and automatically converted it into text using the Google Cloud Speech (v2) platform.[6] We used the same tool to identify the speaker, leveraging its speaker diarization feature.

All transcripts underwent a careful manual proofreading process to correct automatic transcription errors, thus ensuring that the transcripts are perfectly aligned with the source. The transcripts were all formatted in the same way: a plain text file with one sentence per line, each starts with either T (for therapist) or C (for client) as shown in Figure 2.

3.3. MITI and CLEAR Annotation

We uploaded the transcripts into an INCEpTION instance (Klie et al., 2018), an open-source an-

[4] www.alexanderstreet.com

[5] casaa.unm.edu

[6] https://cloud.google.com/python/docs/reference/speech/latest/google.cloud.speech_v2.services.speech

```
T: <utterance>
C: <utterance>
⋮
T: <utterance>
C: <utterance>
```

Figure 2: Session transcript structure (T = Therapist, C = Client).

notation platform, for manual annotation. We established three annotation layers: CLEAR and MITI behavioral codes, which are annotated at the utterance level (one turn in the conversation), and MITI global scores, a session-level annotation layer. MITI includes 10 behavioral codes that are assigned to the majority of therapist utterances, with exceptions for utterances that should not be coded. Such utterances include structure statements, greetings, facilitative statements, previous session content, incomplete thoughts, and off-topic material. Further details and examples of these types are provided in the annotation guidelines (see Section 1.1.1). In our dataset, all non-coded statements received a specific label, "structure statement", ensuring that each therapist utterance had a single assigned label. This new label increased the size of our MITI behavioral label set to 11. Each utterance received only one code. If an annotator wished to assign more than one label to an utterance, it had to be split into multiple segments. Each segment was considered a distinct utterance and was assigned a single label. As a result, a session could have multiple consecutive 'T' or 'C' utterances.

We also assigned the four MITI global scores, each provided for the entire session. The four global scores are determined using a 5-point Likert scale. According to the official MITI guidelines (Section 1.1.1), the annotators should start with a default score of "3" and adjust upward or downward as necessary. Generally speaking, a score of "3" can also indicate mixed practice. The four global scores are:

Cultivating Change Talk: This scale measures the extent to which the clinician actively encourages the client's own language in support of the change goal and their confidence in making that change.

Softening Sustain Talk: This scale measures the extent to which the clinician avoids focusing on reasons against change or for maintaining the status quo.

Partnership: This scale measures the extent to which the clinician conveys an understanding that expertise and wisdom about change primarily reside within the client.

Empathy: This scale measures the extent to which the clinician tries to understand or grasp the client's perspective and experience, essentially attempting to "try on" what the client feels or thinks.

We used CLEAR for annotating client talk. It comprises of two labels: "change-talk" and "counter-change-talk", both of which are exclusively used for labeling client utterances. Per the guidelines, neutral client language is not to be coded. We automatically labeled every non-coded utterance with the special label "neutral".

We trained three annotators (all are undergraduate psychology students) to annotate the CLEAR and MITI coding schemes according to their respective guidelines. The annotators were trained by an expert clinical psychologist, knowledgeable in the field of MI and behavioral coding in MITI and CLEAR. Following two annotation training sessions, all the annotators and the expert psychologist annotated the same fifteen sessions. This was done in order to improve the inter-annotator agreement.

For the MITI behavioral code annotation, we measure inter-annotator agreement between every one of the three annotators and the expert psychologist using Cohen's kappa. The initial pairwise inter-annotator agreement was found to be approximately 0.6 on average. After examining the annotations during an open discussion regarding disagreements, the focus shifted to addressing challenges faced by annotators in discerning between simple and complex reflections during coding. The objective was to improve clarity and consensus in their annotations. In addition, the group set technical guidelines to handle some technical issues we encountered along the annotation process. An example of such a guideline is: "When annotating a sentence, include the punctuation marks in the current behavioral code". The group was thus able to improve the agreement to 0.67 on average after annotating additional five sessions. In order to further improve the agreement between the annotators, three additional sessions were annotated. This resulted in final inter-annotator agreement values of 0.74, 0.84, and 0.68, respectively, when each of the three annotators was compared to the expert psychologist.

Once that agreement level was achieved, each annotator was randomly assigned roughly 60 sessions of varying lengths. These assignments were mutually exclusive, ensuring that each session was annotated individually. The students were paid 19$ per hour for this job. After completing the annotation process, we curated the sessions for which multiple annotations are available, by prioritizing the expert psychologist's annotation; otherwise, we used the assigned annotator's contribution, resulting in a single coded dataset.

We measured inter-annotator agreement for CLEAR as well. CLEAR is less complex than MITI, and as a result, we achieved a kappa of 0.9, averaged over the agreements between every annotator and the expert psychologist.

For the global scores in MITI, we measured the Pearson correlation coefficients due to their basis on a Likert scale; Figure 3 displays the correlation matrix among annotators, for all four types of global score. The average correlation values, when comparing one of the three annotators with the expert psychologist, are 0.64, 0.39, 0.50, and 0.59 for empathy, partnership, softening sustain talk, and cultivating change talk, respectively. These results indicate a strong positive correlation between the annotations in all global scores, with the exception of partnership. All three global scores—empathy, softening sustain talk, and cultivating change talk—have some related MITI and CLEAR behavioral codes (simple and complex reflection are related to empathy, counter-change talk is related to softening sustain talk, and change talk is related to cultivating change talk), which make it easier for the annotators to agree on those scores. Partnership is the only global score which does not have such corresponding labels; hence, we believe it is more challenging to achieve consensus on this score.

Overall, the new dataset comprises 242 MI sessions. Only 236 sessions out of the 242 include global scores, manually assigned by the annotators. The sessions we did not annotate with global scores were too short, falling below the length limit defined by the guidelines. The full descriptive statistics of the dataset is provided in Table 1.

Number of sessions	242
Total utterance count	15,627
Total therapist utterance count	8,389
Total client utterance count	7,238
Average utterance count per session	64
Total word count	339,045
Average word count per session	1,401

Table 1: Descriptive statistics of our new dataset.

Figure 4 displays the MITI code distribution, and the code distribution of CLEAR is summarized in Figure 5; both are unbalanced.

Ethics Statement

Annotating, analyzing, and modeling publicly available therapy data is of great importance as it is the only available alternative to real transcripts of therapy sessions without privacy violations. The study was approved by the Ethics Committee of The Reichman University on May 9th 2023.

4. Evaluation

We explore whether the annotated dataset can be used for automated annotation in the future. We trained models (mostly fine-tuning and "prompt engineering") and carried out several experiments evaluating the quality of automated label classification and global score estimation.

4.1. MITI and CLEAR Codes Classification

We trained a classifier to predict MITI and CLEAR codes, aiming to tag each utterance with the relevant behavioral code. An utterance, when taken out of its context in the session, can be misleading; this is especially relevant to client talk type prediction. Thus, we extended the classification task to include context. We tested three context types:

No-context: Using only the target utterance, without the surrounding text.

One: Concatenating the previous speaker utterance with the target utterance.

Volley: Concatenating the previous therapist and client utterance turnaround.

To predict the codes we fine-tuned the traditional BERT-base-uncased (Devlin et al., 2019) model from the Hugging Face platform (Wolf et al., 2019), for the token-sequence classification task, using the CLS-pooled vector for classification. We individually trained one classifier for therapist utterances to predict MITI codes and another classifier for client utterances to predict CLEAR codes. We split the dataset into training (169 sessions) and test (73 sessions) sets. The results presented in the following section are based on the test set. For both classifiers, we used a standard learning rate of $3\mathrm{e}{-5}$, a weight decay of 0.01, and trained the model for 10 epochs with a batch size of 16. Additionally, to prevent over-fitting, we implemented an early stopping strategy. All utterances fall under the 512-token limit imposed by BERT.

4.2. Global Score Prediction

As mentioned before, there are four global scores assigned on a 5-point Likert scale. The global-score prediction task is designed to take the entire session transcript as input and predict a single value (1-5) for one of the four global scores. It requires potentially processing long transcripts;

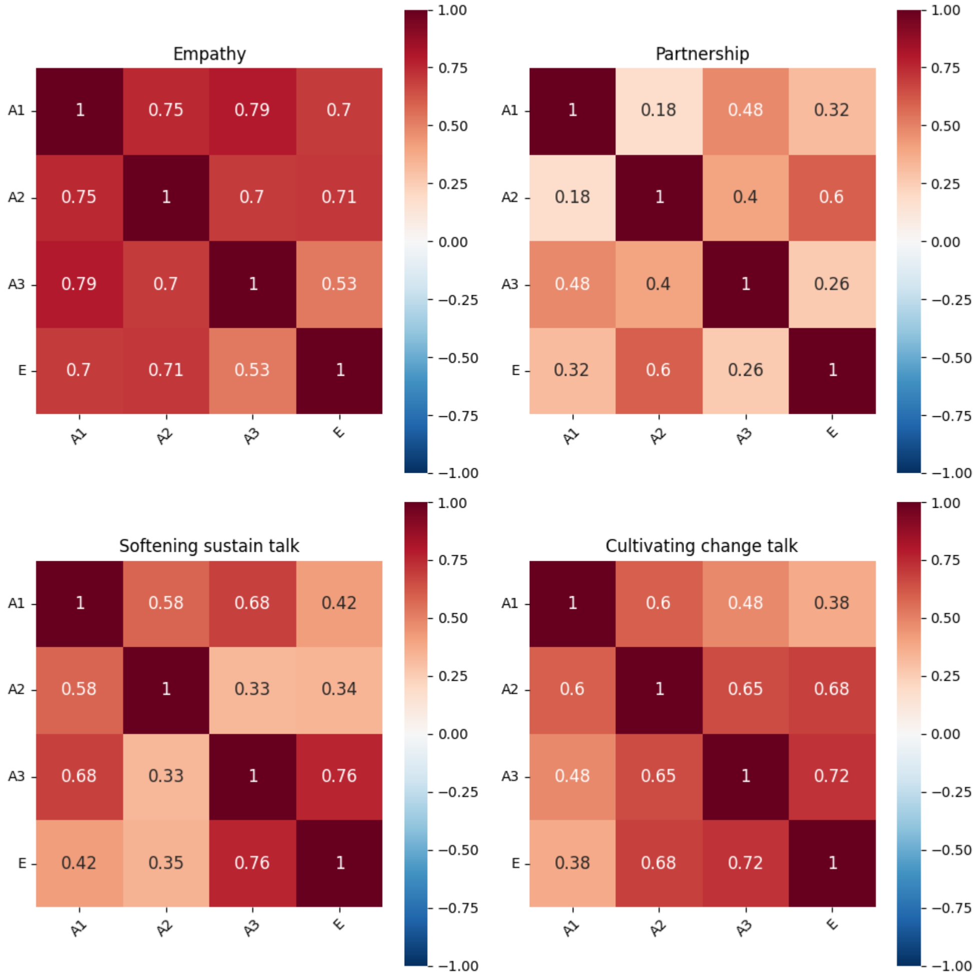

Figure 3: The Pearson correlation values between the annotators for the four MITI global scores. A1 – A3 are the trained annotators and E is the expert annotator.

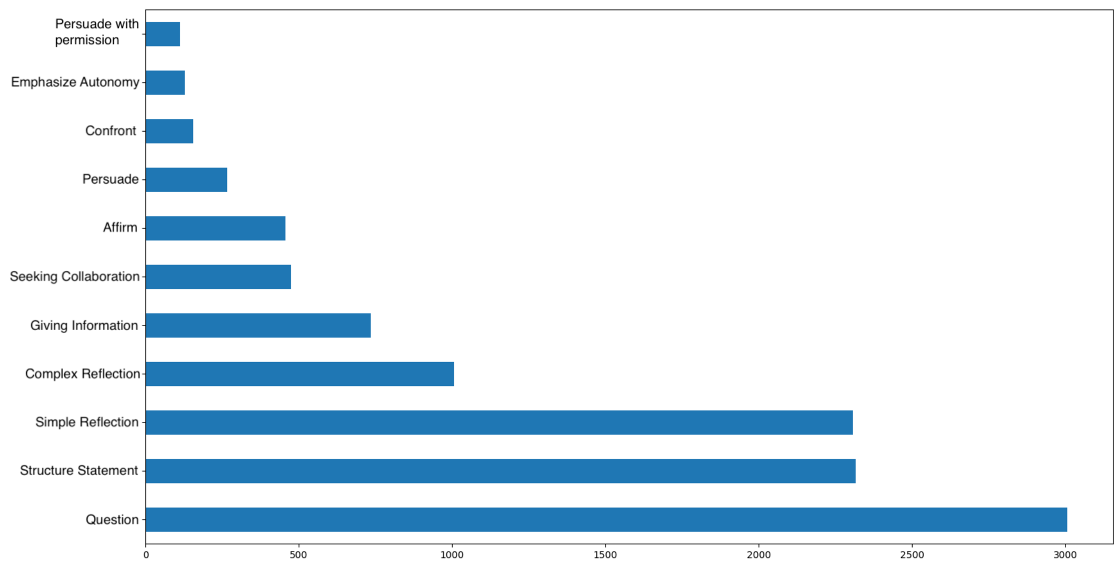

Figure 4: MITI code distribution in our MI-TAGS dataset.

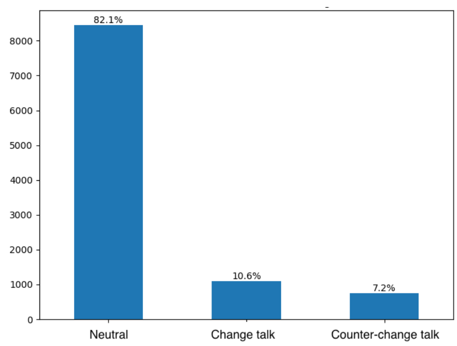

Figure 5: CLEAR label distribution.

handling large context inputs is a well known challenge for NLP models. Therefore, we decided to handle this task with a large language model (LLM), in a completely zero-shot settings using two prompt styles: one that provides the definition of the specific global score, as provided in the MITI guidelines, and another one that uses a short summary of that description. The description was generated automatically using the online version of ChatGPT (GPT-3.5). Due to input length limitations, we removed the score examples provided in the MITI guidelines from both prompts. Each prompt was followed by the entire session transcript, and then by a closing request, asking the model to predict the score. The model usually provided the score embedded within a textual explanation. To extract the numeric score, we employed a simple rule-based logic. The prompt structure is depicted in Figure 6; the full and summarized descriptions of the global scores are provided in Appendix A. The session transcripts are provided either as shown in Figure 2 or enriched with MITI and CLEAR behavioral codes, as shown in Figure 7.

Context Size	Accuracy	Macro F1	ROC AUC
No-context	0.70	0.40	0.90
One	0.70	0.42	0.91
Volley	0.70	0.41	0.91

Table 2: Predicting MITI behavioral codes

To predict each of the four global scores we tested four LLMs: two from OpenAI, `gpt-3.5-turbo-0613` and `gpt-4-0613`, one open-source model, Llama-2 (`meta-llama/Llama-2-70b-chat-hf`), and PaLM-2 (`text-bison-001`). Due to each LLM's individual token input limitations, we assessed their performance on 47 sessions that fit within these limits and were annotated by the expert psychologist.

```
You are a motivational interviewing
assistant tasked with evaluating the
level of [Empathy] demonstrated by a
therapist in a dialogue transcript.

<[Empathy] global score description>

<Session transcript>

Assign a score (1-5) for [Empathy]
and provide a step by step explana-
tion, highlighting specific behav-
iors justifying the assigned score.
```

Figure 6: For predicting global scores, we use a prompt structure. The [Empathy] component is substituted with other global scores for individual predictions. The <> symbol acts as a placeholder, which we populate with the relevant text. The full and summarized global score descriptions are shown in Appendix A.

```
T: [Complex Reflection]<utterance>
C: [Change Talk] <utterance>
⋮
T: [Structure Statement] <utterance>
C: [Neutral] <utterance>
```

Figure 7: Session transcript enriched with MITI and CLEAR behavioral codes. ({T}herapist, {C}lient).

5. Results

We present the MITI behavioral code classification results in Table 2. Our results are significantly higher than the corresponding chance level results (accuracy = 0.18, Macro F1 = 0.09). Additionally, it is evident that the context size variations do not result in significant differences.

Table 3 provides a detailed description of code prediction for each label, complemented by Figure 8 which includes the chance level F1 scores per label. Overall, we see good F1 scores for most labels with large support.

During the annotation project our annotation team noticed that certain codes, such as simple vs complex reflection, posed a challenge to our human annotators as mentioned in Section 3. Inspecting the confusion matrix (see precision and recall scores in Table 3) shows that these same codes posed a challenge for the models as well.

Table 4 summarizes the classification performance of the CLEAR classifier. Similar to MITI, we see no significant difference in performance between the different context sizes, although it does seem that the classifier may benefit from a longer context. Table 5 displays the label-specific perfor-

Class	Precision	Recall	F1-Score	Support
Structure Statement	0.84	0.80	0.82	505
Question	0.87	0.91	0.89	730
Complex Reflection	0.45	0.22	0.30	228
Simple Reflection	0.61	0.77	0.68	563
Seeking Collaboration	0.33	0.33	0.33	93
Giving Information	0.66	0.74	0.70	172
Emphasize Autonomy	1.00	0.04	0.08	25
Persuade	0.22	0.15	0.18	66
Confront	0.00	0.00	0.00	18
Affirm	0.47	0.58	0.52	93
Persuade with Permission	0.00	0.00	0.00	29

Table 3: MITI classification results per code.

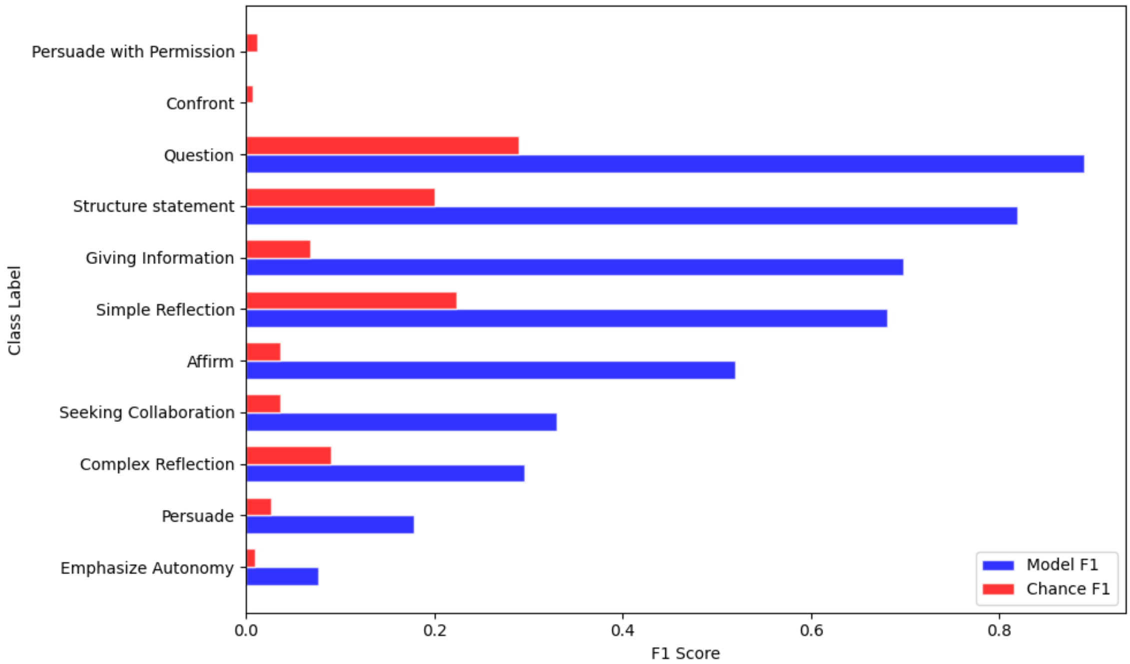

Figure 8: MITI codes accuracy per code: model (Blue) vs. chance level (Red) F1 scores.

mance results.

Context Size	Accuracy	Macro F1	ROC AUC
No-context	0.70	0.69	0.86
One	0.71	0.71	0.87
Volley	0.72	0.72	0.87

Table 4: Predicting CLEAR behavioral codes.

Class	Precision	Recall	F1-Score	Support
CT	0.75	0.74	0.75	228
CCT	0.62	0.71	0.66	150
Neutral	0.77	0.71	0.74	198

Table 5: CLEAR classification results per label (using the best classifier from Table 4).

To assess the global score prediction, we compute the Pearson correlation between the model's predictions and the expert psychologist's annotations. Tables 6-9 report the performance of the best three configuration settings for every global score. Each setting defines the LLM, the type of the global-score description, and whether the transcript comes with annotated labels, as detailed in Section 4.2. Performance is measured using the Pearson correlation between the model's predictions and the annotations from the expert psychologist. Excluding empathy, the results are not significantly different among the top three settings. All correlation scores fall within the range of $[0.34 - 0.63]$, which can be interpreted as ranging from "fair" to "strong" correlation.

When comparing the results of "Tagged text" with "Only text" inputs, one can observe variations in model performance across the various global dimensions of MI While some models excel with raw text inputs for predicting empathy scores, their performance diminishes when behavioral tags are incorporated, suggesting that for certain aspects, such as empathy, the additional contextual information may not significantly enhance predictive accuracy. Conversely, for "Cultivating Change Talk", models generally perform better with "Tagged text", indicating that the contextual nuances captured by the behavioral tags of MITI and CLEAR mayy assist in understanding these behaviors. Similarly, comparing "Summarized" and "Full" prompts reveals some difference in model performance, with models generally exhibiting slightly better performance with summarized prompts. More research is needed to explore optimal configurations for real-life applications. Ultimately, there is no single "winner" configuration for each global score or even a "winner" configuration for each model, highlighting the complexity and variability inherent in MI analysis.

Model	Prompt Type	Input Type	Pearson
GPT-4	Summarized	Only Text	0.63
GPT-4	Full	Tagged Text	0.56
GPT-3.5	Summarized	Only Text	0.56

Table 6: Empathy prediction.

Model	Prompt Type	Input Type	Pearson
GPT-4	Summarized	Tagged Text	0.46
GPT-4	Full	Only Text	0.45
Llama 2	Summarized	Tagged Text	0.44

Table 7: Partnership prediction.

Model	Prompt Type	Input Type	Pearson
PaLM 2	Full	Tagged Text	0.44
Llama 2	Summarized	Only Text	0.42
GPT-4	Full	Tagged Text	0.41

Table 8: Cultivating Change Talk prediction.

Model	Prompt Type	Input Type	Pearson
GPT-4	Full	Only Text	0.36
GPT-3.5	Full	Only Text	0.36
GPT-4	Summarized	Only Text	0.34

Table 9: Softening Sustain Talk prediction.

6. Conclusion and Future Work

In this study, we introduce a new dataset comprising manually transcribed and professionally annotated MI counseling dialogues. We detail the distribution of therapist and client utterances, behavioral tags, and global scoring at the session level, leveraging the well-established coding systems, MITI and CLEAR.

We provide evidence that carefully prompted or fine tuned LLMs can be used for automatically tagging utterances with MITI and CLEAR codes, as well as automatically assigning the MITI global scores to the entire session according to the four MITI categories. We are releasing the dataset for the broader research community at https://advanced-reality-lab.github.io/MI-TAGS/.

As we witness further advancements in LLMs, the demand for more data will grow, especially for lesser-explored tasks like MI. Rather than manually coding MI sessions, MI experts can focus on refining prompts, thus enhancing automated classification on a broader scale. This iterative approach has the potential to continually enhance the formulation of psychological expertise. Practically, this strategy represents a fusion of "statistical AI", fundamental to LLMs, and "traditional" knowledge-based AI.

Future work could potentially focus on follow-up challenging tasks, such as better distinguishing between simple and complex reflection codes.

With the introduction of our dataset and models, we hope to facilitate further research towards creating both offline and real-time evaluation tools for clinicians and contribute to the advancement of automated therapeutic interventions.

7. Acknowledgments

This work was partially supported by European Commission Horizon 2020 projects GuestXR (101017884) and Socrates (951930).

8. Bibliographical References

Roger Bakeman and Stephen Haynes. 2015. *Behavioral Observation*.

Jie Cao, Michael Tanana, Zac E. Imel, Eric Poitras, David C. Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikolaos Flemotomos, Victor R. Martinez, Zhuohao Chen, Karan Singla, Victor Ardulov, Raghuveer Peri, Derek D. Caperton, James Gibson, Michael J. Tanana, Panayiotis G. Georgiou, Jake Van Epps, Sarah P. Lord, Tad Hirsch, Zac E. Imel, David C. Atkins, and Shrikanth Narayanan. 2021. "Am I a good therapist?" automated evaluation of psychotherapy skills using speech and language technologies. *CoRR*, abs/2102.11265.

Paul J Hershberger, Yong Pei, Dean A Bricker, Timothy N Crawford, Ashutosh Shivakumar, Miteshkumar Vasoya, Raveendra Medaramitta, Maria Rechtin, Aishwarya Bositty, and Josephine F Wilson. 2021. Advancing motivational interviewing training with artificial intelligence: ReadMI. *Adv. Med. Educ. Pract.*, 12:613–618.

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

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Aniko Maraz, Eva Katzinger, and Sunghwan Yi. 2021. Potentially addictive behaviours increase during the first six months of the Covid-19 pandemic. *Journal of Behavioral Addictions*, 10.

William R Miller and Stephen Rollnick. 1993. [book review] motivational interviewing, preparing people to change addictive behavior. *Journal of Studies on Alcohol*, 54:507–507.

Theresa Moyers, Tim Martin, Jennifer Manuel, Stacey Hendrickson, and William Miller. 2005. Assessing competence in the use of motivational interviewing. *Journal of substance abuse treatment*, 28:19–26.

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. 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)*, pages 1426–1435, Vancouver, Canada. 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.

S. Rollnick, W.R. Miller, and C. Butler. 2008. *Motivational Interviewing in Health Care: Helping Patients Change Behavior*. Applications of motivational interviewing. Guilford Publications.

Leili Tavabi, Trang Tran, Kalin Stefanov, Brian Borsari, Joshua Woolley, Stefan Scherer, and Mohammad Soleymani. 2021. Analysis of behavior classification in motivational interviewing. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 110–115, Online. Association for Computational Linguistics.

Tracey Wade, Anna Frayne, Sally-Anne Edwards, Therese Robertson, and Peter Gilchrist. 2009. Motivational change in an inpatient anorexia nervosa population and implications for treatment. *The Australian and New Zealand journal of psychiatry*, 43:235–43.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

9. Language Resource References

Pérez-Rosas, Verónica and Wu, Xinyi and Resnicow, Kenneth and Mihalcea, Rada. 2019. *What Makes a Good Counselor? Learning to Distinguish between High-quality and Low-quality Counseling Conversations*. Association for Computational Linguistics.

Welivita, Anuradha and Pu, Pearl. 2022. *Curating a Large-Scale Motivational Interviewing Dataset*

Using Peer Support Forums. International Committee on Computational Linguistics.

Wu, Zixiu and Balloccu, Simone and Kumar, Vivek and Helaoui, Rim and Reiter, Ehud and Recupero, Diego Reforgiato and Riboni, Daniele. 2022. *Anno-MI: A dataset of expert-annotated counselling dialogues*. IEEE.

A. Appendix: Descriptions Used for Global Score Prediction

Figures 9-12 provide the full and summarized descriptions of each of the four global scores.

B. Appendix: LLM Response

Figure 13 shows the GPT-4 response for predicting the empathy global score.

C. Appendix: Examples for Annotator Disagreement

Table 10 provides some examples for disagreements among the annotators.

Empathy, full description.
Empathy refers to the extent to which the clinician actively works to understand the client's perspective and experiences, striving to comprehend their emotions, thoughts, and unspoken meanings. Please utilize the following scale to assess the therapist's behavior in the dialogue: 1: Reflects Clinician gives little or no attention to the client's perspective. 2: Reflects Clinician makes sporadic efforts to explore the client's perspective. Clinician's understanding may be inaccurate or may detract from the client's true meaning. 3: Reflects Clinician is actively trying to understand the client's perspective, with modest success. 4: Reflects Clinician makes active and repeated efforts to understand the client's point of view. Shows evidence of accurate understanding of the client's worldview, although mostly limited to explicit content. 5: Reflects Clinician shows evidence of deep understanding of the client's point of view, not just for what has been explicitly stated but also for what the client means but has not yet said. It's important to keep in mind that the objective of this evaluation is to assess the extent to which the therapist demonstrates understanding and attempts to grasp the client's feelings, thoughts, and perspectives. Empathy should not be confused with other positive qualities like sympathy or support. Reflective listening is a component of empathy, but this rating is designed to capture the therapist's overall efforts to understand and convey that understanding to the client. Higher ratings on the Empathy scale should be given when the therapist exhibits accurate comprehension of the client's worldview and emotions, even beyond the explicit content. Lower scores should be assigned when the therapist appears disinterested in the client's viewpoint.

Empathy, summarized description.
Empathy reflects the therapist's ability to understand, validate and demonstrate empathy towards the client's thoughts and feelings. Assess the empathy score based on the following behaviors: complex reflections, simple reflections and affirmations. Consider the overall tone, language, and non-judgmental attitude exhibited by the interviewer.

Figure 9: Empathy descriptions.

Utterance	Annotator Label	Expert Label
"You want more out of life"	Simple Reflection	Complex Reflection
"So, a few drinks is nothing new to you"	Complex Reflection	Simple Reflection
"Just wanted to check in and see how things are going"	Question	Structure Statement
"Yeah, I mean you're healthy except for this problem"	Simple Reflection	Giving Information

Table 10: Disagreements among annotators.

Softening Sustain Talk, full description.
Softening Sustain Talk refers to the extent to which the clinician avoids focusing on the client's reasons for maintaining the status quo and instead aims to shift the conversation towards building motivation for change. Please use the following scale to assess the therapist's behavior in the dialogue: 1: Reflects Clinician consistently responds to the client's language in a manner that facilitates the frequency or depth of arguments in favor of the status quo. 2: Reflects Clinician usually chooses to explore, focus on, or respond to the client's language in favor of the status quo. 3: Reflects Clinician gives preference to the client's language in favor of the status quo, but may show some instances of shifting the focus away from sustain talk. 4: Reflects Clinician typically avoids an emphasis on client language favoring the status quo. 5: Reflects Clinician shows a marked and consistent effort to decrease the depth, strength, or momentum of the client's language in favor of the status quo. Keep in mind that the aim of this evaluation is to determine the extent to which the therapist avoids dwelling on reasons against change and instead employs motivational interviewing techniques to encourage motivation for change. High scores should be awarded when the therapist effectively navigates away from sustain talk and low scores when they spend considerable time discussing barriers to change, even if using MI-consistent techniques. Remember that the absence of sustain talk engagement is also a factor to consider for high scores if the clinician does not actively evoke it.

Softening Sustain Talk, summarized description.
Softening Sustain Talk reflects the therapist's ability to address the client's expressions of resistance or reluctance to change in a supportive manner, fostering openness to the change process. Assess the Softening Sustain Talk score based on the following: Assign higher scores when the interviewer skillfully steers away from sustain talk and towards cultivating motivation for change. Take note that lower scores are appropriate when the interviewer dedicates considerable time to discussing barriers to change, even if they use motivational interviewing-consistent techniques. Remember that achieving high scores also involves recognizing situations where the clinician refrains from actively evoking sustain talk. Evaluate the extent to which the interviewer navigates conversations away from sustain talk while employing motivational interviewing strategies.

Figure 10: Softening sustain talk descriptions.

Cultivating Change Talk, full description.

Cultivating Change Talk refers to the extent to which the clinician actively encourages and reinforces the client's language favoring the desired change goal and their confidence in achieving that change. Please utilize the following scale to assess the therapist's behavior in the dialogue:1: Reflects Clinician shows no explicit attention to, or preference for, the client's language in favor of changing.2: Reflects Clinician sporadically attends to client language in favor of change - frequently misses opportunities to encourage change talk.3: Reflects Clinician often attends to the client's language in favor of change, but misses some opportunities to encourage change talk.4: Reflects Clinician consistently attends to the client's language about change and makes efforts to encourage it.5: Reflects Clinician shows a marked and consistent effort to increase the depth, strength, or momentum of the client's language in favor of change.in mind that the objective of this evaluation is to measure how actively the therapist supports the client's expressions of readiness for change and their commitment to the change goal. Higher ratings on the Cultivating Change Talk scale should be given when the therapist consistently emphasizes and encourages change talk, fostering a dialogue focused on the desired change. Lower scores should be assigned when the therapist misses opportunities to engage with change talk or prioritizes other aspects of the interaction. It's important not to penalize clinicians if clients are not forthcoming with change talk or don't respond positively to change-evoking efforts. Remember, interactions low in Cultivating Change Talk can still be empathic and clinically valid.

Cultivating Change Talk, summarized description.

Cultivating Change Talk represents the therapist's skill in eliciting and reinforcing the client's expressions of motivation and commitment to change. Assess the Cultivating Change Talk score based on the following behaviors: Questions, Emphasize Autonomy and Persuade with Permission. Consider the overall effectiveness of the interviewer in cultivating the client's expressions of motivation and commitment to change.

Figure 11: Cultivating change talk descriptions.

Partnership, full description.
Partnership refers to the degree to which the clinician actively engages in collaboration and power sharing with the client, acknowledging the client's expertise and wisdom in their journey towards change. Please make use of the following scale to assess the therapist's behavior in the dialogue: 1: Reflects Clinician actively assumes the expert role for the majority of the interaction with the client. Collaboration or partnership is absent. 2: Reflects Clinician superficially responds to opportunities to collaborate. 3: Reflects Clinician incorporates the client's contributions but does so in a lukewarm or erratic fashion. 4: Reflects Clinician fosters collaboration and power sharing so that the client's contributions impact the session in ways that they otherwise would not. 5: Reflects Clinician actively fosters and encourages power sharing in the interaction in such a way that the client's contributions substantially influence the nature of the session. It's essential to keep in mind that the goal of this evaluation is to determine how effectively the therapist conveys an understanding that the client holds the primary expertise and wisdom regarding their change process. Higher ratings on the Partnership scale should be given when the therapist actively fosters a collaborative and equal partnership with the client, acknowledging the client's valuable insights. Conversely, lower scores should be assigned when the therapist assumes an expert role or does not actively encourage collaboration. Remember that clinicians who score low on this scale may exert significant influence over the interaction.

Partnership, summarized description.
Partnership reflects the collaborative and respectful nature of the interaction between the therapist and the client. Assess the level of Partnership based on the following behaviors: Seeking Collaboration, Emphasize Autonomy and Structure Statements. Consider the overall level of collaboration and mutual respect in the dialogue.

Figure 12: Partnership descriptions.

```
Empathy Score: 4.5
Explanation: 1. Complex Reflections: The therapist consistently utilizes
complex reflections to represent a deep understanding of the client's feel-
ings and experiences. For example, "It seems like the idea that you might
be experiencing signs of depression is not really fitting with how you see
things" and "You have done such a good job, focusing on your kids, so much
for so many years I mean, maybe it feels odd or strange to focus on your-
self" demonstrate that the therapist is actively seeking to understand the
client's perspective and make sense of her experiences. 2. Simple reflec-
tions: The therapist also uses simple reflections to show that he is lis-
tening and understanding the client's feelings, such as when he acknowl-
edges the overwhelming nature of the situation, "It must be overwhelming
to think about this." 3. Affirmations: The therapist recognizes and af-
firms the client's efforts and positive attributes, for example saying, "I
mean you put the needs of your children. First, I mean it sounds like you,
don't even have time for anything else. I respect that" and "I don't think
you are damaging your kids. I mean the fact that you're even worried about
this shows what a really caring mom you are." These affirmations acknowl-
edge the client's hard work and show a great deal of respect and empathetic
understanding. 4. Overall tone, language, and non-judgmental attitude:
The therapist exhibits a warm, caring, and non-judgmental attitude towards
the client. He uses open-ended questions ("Do you want to know more...?"),
invites the client to share her own feelings and thoughts, and provides in-
formation in a respectful and non-threatening way. The score is not a full
5 because there could still be some improvements made. For instance, the
therapist could exhibit more empathy by further exploring the client's feel-
ings about her stress over not having time or her worries about her chil-
dren, rather than jumping right into providing additional information about
depression.
```

Figure 13: GPT-4 response for predicting the global empathy score.