

# PHASEMI: A MOTIVATIONAL INTERVIEWING DATASET FOR ENHANCING PHASE PROGRESSION IN LLM-BASED COUNSELING

Jina Kim<sup>1,2\*</sup>, Myeongho Jeon<sup>2\*</sup>, Soohyun Cho<sup>3</sup>, Chae-Gyun Lim<sup>2</sup>,  
Jongmin Lim<sup>4</sup>, Haewon Min<sup>5</sup>, Eunho Yang<sup>2†</sup>

<sup>1</sup>KT Corporation <sup>2</sup>KAIST <sup>3</sup>Keimyung University  
<sup>4</sup>Korea Army Academy <sup>5</sup>Korea University

## Abstract

The growing demand for scalable mental health support has increased interest in AI-based counseling systems grounded in Motivational Interviewing (MI). However, existing MI datasets do not explicitly model the structured progression of MI phases, which is essential for effective and goal-oriented counseling. To address this gap, we introduce PHASEMI, a phase-structured MI dataset, together with a data generation framework that employs therapist, client, and supervisor LLMs to explicitly control phase transitions. Compared to the best alternative baseline, PHASEMI achieves improved coverage of MI phases, with gains of 12.3% in Exploring, 37.6% in Guiding, and 61.1% in Choosing, and experimental evaluations demonstrate that it yields higher overall counseling quality than baseline datasets.

## 1 Introduction

**Important:** We explore how large language models support digital MI, but they should NOT replace professional psychological treatment.

Mental health challenges are increasingly prevalent, placing substantial strain on existing health-care systems. This growing burden has led to a rising demand for scalable mental health support solutions (Pedrelli et al., 2015). In response, AI-based psychological counseling chatbots are gaining attention as easy-to-access, low-cost tools that people can use without worrying about stigma, either as alternatives to or alongside traditional therapy (Boucher et al., 2021). For such systems to be effective and trustworthy, they must be grounded in counseling approaches that have been systematically validated in clinical practice, since the conversational agents are reported to be capable of both amplifying and violating a therapeutic process (Fitzpatrick et al., 2017).

\*Equal contribution.

†Correspondence to: eunhoy@kaist.ac.kr

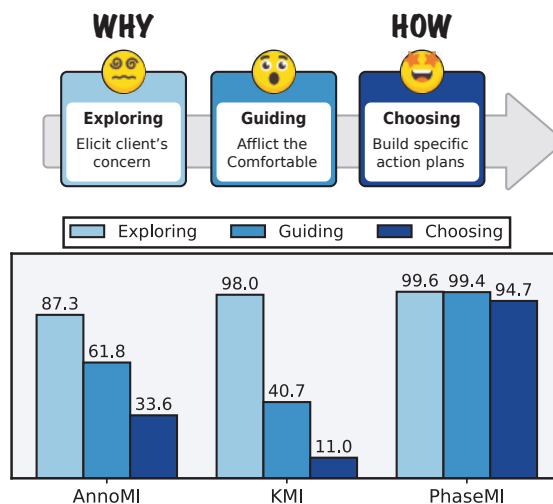


Figure 1: MI phase coverage. While existing datasets (AnnoMI, KMI) primarily focus on the Exploring phase, PHASEMI provides coverage across all MI phases.

As a counseling theory, Motivational Interviewing (MI) is a client-centered, goal-oriented counseling approach that aims to elicit and strengthen individuals' intrinsic motivation for change by exploring and resolving ambivalence. MI has demonstrated effectiveness in supporting behavioral change and personal growth among individuals experiencing uncertainty about change. However, the development of MI-aligned AI systems is limited by the lack of large, high-quality datasets due to privacy restrictions.

To this end, AnnoMI (Wu et al., 2022, 2023c), an expert-annotated MI dialogue dataset, was introduced. However, it is limited in scale (*i.e.*, 133 samples), and some sessions are partially truncated due to privacy concerns, which may restrict its effectiveness for training large language models (LLMs). KMI (Kim et al., 2025a), a synthetic MI dataset consisting of 1,000 Korean dialogues generated using LLMs, addresses scalability issues. Nevertheless, **neither dataset explicitly**

**accounts for coverage of the core counseling session phases**—Exploring, Guiding, and Choosing—which are designed to progressively build client motivation and transition toward action planning (Figure 1).

To address this issue, we propose a new dataset, **Phase-Structured Motivational Interviewing Dataset (PHASEMI)**<sup>1</sup>, which comprises 13,897 dialogues, along with a corresponding data generation framework that is explicitly designed to support phase-structured MI. Accordingly, we employ three LLMs that assume the roles of a therapist, a client, and a supervisor. Within the proposed three-LLM framework, a supervisory model controls MI phase transitions by assessing whether the dialogue should remain in the current phase or progress to a subsequent one. This design enables the generation of high-quality counseling data with consistent, accurate alignment with the MI phase structure.

Experimental results demonstrate the effectiveness of our phase-structured design. PHASEMI outperforms existing MI datasets in counseling quality across multiple evaluation criteria and achieves substantially better coverage of all MI phases, particularly the later Guiding and Choosing stages. Moreover, models fine-tuned on PHASEMI show improved performance in counseling quality along with more accurate and complete MI phase progression. These results indicate that PHASEMI enables LLMs to more effectively guide clients toward meaningful behavioral change.

## 2 Preliminaries

### 2.1 Motivational Interviewing

MI is an evidence-based, client-centered counseling intervention that plays a central role in modern counseling practice due to its broad applicability and robust effectiveness across diverse populations (Miller and Rollnick, 2012; Resnicow and McMaster, 2012). Rather than targeting specific disorders, MI focuses on resolving ambivalence and strengthening clients’ intrinsic motivation, making it particularly effective across a wide range of counseling contexts. Empirical studies show that MI is significantly effective for problems such as substance use (*e.g.*, alcohol, marijuana, tobacco, and other drugs), reduction of risky behaviors, and improved engagement in treatment (Lun-

dahl and Burke, 2009). These properties underscore MI’s importance as a general-purpose counseling framework grounded in fundamental mechanisms of human behavior change, rather than techniques tailored to narrow or domain-specific problems (Miller and Rose, 2009).

### 2.2 MI Phase

MI phases refer to a counseling process that moves from understanding the client’s perspective, to eliciting self-motivated reasons for change, and finally to supporting autonomous decision-making (Resnicow and McMaster, 2012). This phased structure is fundamental to MI because it ensures that interventions are temporally aligned with a client’s readiness for change, preventing premature action planning and preserving client autonomy. By progressively building motivation and commitment, the MI phases provide a principled framework for guiding counseling interactions in a way that supports sustained behavior change. Specifically, each phase is defined as follows:

- **Exploring:** The initial stage of the session that counselors elicit the client’s story, build rapport, and obtain a behavioral history, including prior attempts to change. It also includes the collaborative decision with the client of what behaviors to address during the session.
- **Guiding:** The second phase that the counselor may “afflict the comfortable” by moving the conversation toward the possibility of change. The counselor might elicit change talk by asking the client to consider life with and without change, taking into account the client’s broader life goals and values.
- **Choosing:** The last stage where action-oriented approaches are brought to bear. This phase aims to help clients identify a goal, build an action plan, anticipate barriers, and agree on a plan for monitoring, if applicable.
- **Terminating<sup>2</sup>:** An additional stage that summarizes the content of the session and schedules the next follow-up session. This is grounded in one of the MI communication skills, *summary*, which collects reflections and recounts several things the counselor has heard from the client.

<sup>2</sup>This phase aims to describe a natural termination of dialogue, allowing the utterance generators in the Choosing phase to better focus on action planning.

<sup>1</sup>Data and code: [https://github.com/regina-kim/phase\\_mi](https://github.com/regina-kim/phase_mi).

Despite the demonstrated effectiveness of MI-phase-aligned counseling (James et al., 2020), *existing counseling datasets have focused on MI concepts without explicitly modeling MI phases.*

### 3 Related Work

#### 3.1 Domain-specific Tuning

Building a domain-adaptive model begins with a general-purpose instruction–response model that has already been pre-trained on large-scale text corpora and instruction-tuned to follow human intent and task specifications. The model is then further adapted through domain-specific fine-tuning on carefully curated in-domain datasets, enabling it to acquire specialized terminology, domain-specific knowledge, and characteristic interaction patterns. Across diverse domains, a variety of datasets have been introduced to support this adaptation process.

In mathematical reasoning, datasets emphasized multi-step symbolic manipulation and formal logical inference, encouraging structured problem solving beyond surface-level pattern recognition (Hendrycks et al., 2021; Yu et al., 2024; He et al., 2025). Financial-domain datasets focused on numerically grounded reasoning and domain-specific language drawn from real-world financial documents, reflecting practical analysis and decision-making scenarios (Wu et al., 2023a; Yang et al., 2023). Biomedical datasets targeted complex scientific and clinical reasoning, requiring accurate interpretation of medical literature and evidence-based question answering (Jin et al., 2019, 2021). In materials science, large-scale curated resources supported reasoning about chemical structures and material properties, enabling knowledge-intensive scientific inference (Kim et al., 2025b). Finally, legal-domain datasets addressed statutory interpretation and case-based reasoning, reflecting the precision and logical rigor of legal analysis (Guha et al., 2023).

#### 3.2 Counseling Datasets

In response to growing interest in AI-assisted mental health support systems, a wide range of datasets has been introduced across multiple modalities. Audio data captured vocal cues relevant to emotional and psychological states (Gideon et al., 2016; Chang et al., 2011). Video data modeled nonverbal behaviors such as facial expressions and body language (Jiang et al., 2020; Abbas et al., 2021). MRI-based neuroimaging datasets characterized

structural and functional brain patterns associated with mental health conditions (Wang et al., 2013). Electroencephalography (EEG) data provided temporally resolved neural signals related to cognition and emotion (Wang et al., 2016). Multi-modal counseling datasets integrating text, audio, and video enabled holistic modeling of therapeutic interactions (Zhu et al., 2023).

Several datasets were introduced to support mental health analysis that incorporates textual representations. WorryWords (Mohammad, 2024) provided manually curated anxiety associations for over 44,450 English words, enabling fine-grained lexical analysis of anxiety-related language. MMPsy (Qin et al., 2025) was a large-scale multi-modal psychological assessment corpus containing audio recordings and transcripts from Mandarin-speaking adolescents involved in automated anxiety and depression assessments. In addition, Aich et al. (2022) released manually transcribed interactions with individuals clinically diagnosed with bipolar disorder and schizophrenia, supporting clinical mental health dialogue research. To address scalability and privacy challenges in mental health AI research, synthetic data generation emerged as a key strategy, with common approaches using LLMs to generate multi-turn dialogues (Wu et al., 2023b; Qiu et al., 2024) or question–answer pairs (Xu et al., 2025; Liu et al., 2023) derived from real data.

MI counseling theory was further leveraged to develop specialized counseling datasets. AnnoMI (Wu et al., 2022, 2023c) was an expert-annotated MI dataset consisting of 133 counseling dialogues. Building on this line of work, Kim et al. (2025a) introduced a framework for simulating MI sessions by modeling therapists’ behavioral decisions with an MI forecaster and generating utterances using large language models (LLMs). Using this framework, they presented KMI, a theory-grounded synthetic dataset of 1,000 Korean MI dialogues that was validated through expert evaluation and MI-based metrics. However, *neither dataset explicitly accounted for coverage of the core MI counseling session phases.*

### 4 PHASEMI: A Phase-Structured Motivational Interviewing Dataset

To explicitly model MI phases, we propose a new data generation framework implemented using three LLMs—a therapist, a client, and a supervisor (Section 4.1). Using this framework, we introduce PHASEMI, a phase-structured MI dataset

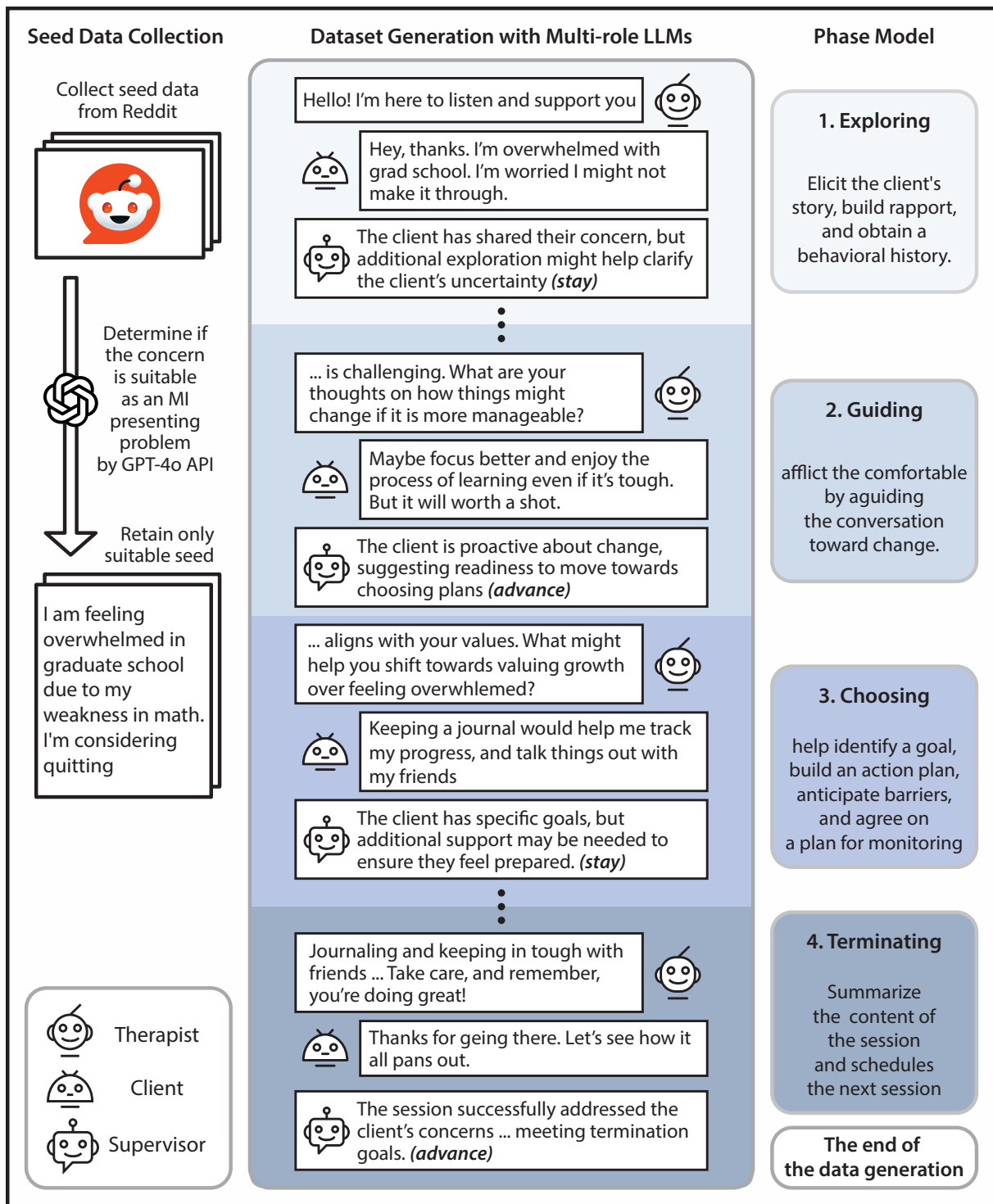


Figure 2: Dataset generation with multi-role LLMs. Seed data from Reddit were filtered, then therapist and client LLMs generated dialogues under a supervisor LLM that explicitly controlled counseling phase progression.

(Section 4.2).

#### 4.1 Generating Framework

**Seed data collection.** We first collect seed data from Reddit<sup>3</sup> filtering them with following taxonomy: (i) academic concerns, (ii) interpersonal concerns, (iii) emotional concerns, and (iv) career

<sup>3</sup>College, r\gradSchool for 2022-2024.

concerns.

We use the GPT-4o API to determine whether a client's concern is suitable as an MI presenting problem<sup>4</sup>. Concerns that match this criterion are

<sup>4</sup>A presenting problem refers to the issue offered by a client as the primary reason for seeking treatment (e.g., depression, anxiety, or family or marital problems), which may become the focus of treatment (VandenBos, 2007).

retained, while those that do not are excluded from the dataset. The detailed information about taxonomy is described in Section B. These concern-focused seed samples are then used to construct corresponding counseling dialogues through the following procedure.

**Dataset generation with multi-role LLMs.** Naively using LLMs to simulate both the therapist and the client does not guarantee control over counseling phases, as such models lack explicit awareness of phase structure. To address this, we generate the dataset using three LLMs that assume the roles of a *therapist*, a *client*, and a *supervisor*. Dialogue generation proceeds in a turn-based manner with explicit control over counseling phases.

Let  $t$  denote the dialogue turn index. The conversation history up to turn  $t - 1$  is denoted by  $H_t = \{(u_1^T, u_1^C), \dots, (u_{t-1}^T, u_{t-1}^C)\}$ , which consists solely of therapist and client utterances, excluding any supervisor guidance. The current counseling phase is denoted by  $p_t \in \mathcal{P}$ , where  $\mathcal{P}$  is the set of predefined phases (including Exploring, Guiding, Choosing, and Terminating). Each phase  $p_t$  is associated with a textual definition  $D(p_t)$  and a set of counseling skills  $S(p_t)$  that are relevant to that phase. The concern seed data describing the client’s initial concerns is denoted by  $c$ .

At turn  $t$ , the therapist generates a counselor utterance  $u_t^T$  conditioned on the conversation history, the current phase information, and a generative instruction:

$$u_t^T \sim \mathcal{M}_{therapist}(u \mid H_t, p_t, D(p_t), S(p_t)), \quad (1)$$

where  $\mathcal{M}_{therapist}$  denotes the therapist LLM. Conditioning on both the phase definition and the associated skill set encourages phase-appropriate and skill-aware counselor responses.

Given the therapist’s utterance, the client produces a response  $u_t^C$  conditioned on the updated conversation history and the concern seed data:

$$u_t^C \sim \mathcal{M}_{client}(u \mid H_t \cup \{u_t^T\}, c), \quad (2)$$

where  $\mathcal{M}_{client}$  denotes the client LLM. This formulation ensures that the client’s responses remain consistent with the predefined concern context.

Following each therapist–client exchange, the supervisor determines whether the dialogue should remain in the current phase or transition to the next phase. The supervisor has access to all available information, including the conversation history, the

current phase information, and the concern seed data:

$$a_t \sim \mathcal{M}_{supervisor}(a \mid H_t \cup \{u_t^T, u_t^C\}, p_t, c), \quad (3)$$

where  $a_t \in \{\text{stay}, \text{advance}\}$  and  $\mathcal{M}_{supervisor}$  denotes the supervisor LLM. The phase is updated according to

$$p_{t+1} = \begin{cases} p_t, & \text{if } a_t = \text{stay}, \\ \text{next}(p_t), & \text{if } a_t = \text{advance}. \end{cases} \quad (4)$$

If the current phase  $p_t$  is a terminal phase and the supervisor selects advance, dialogue generation is terminated.

## 4.2 PHASEMI

As a result, a dialogue instance is represented as a sequence of therapist–client utterance pairs with associated phase annotations. The final dataset is defined as

$$\mathcal{D} = \left\{ \left( \{(u_t^T, u_t^C)\}_{t=1}^{T_d}, \{p_t\}_{t=1}^{T_d} \right) \right\}_{d=1}^N, \quad (5)$$

where  $T_d$  denotes the number of turns in dialogue  $d$ ,  $p_t$  is the counseling phase at turn  $t$ , and  $N$  is the total number of generated dialogues. Example single-turn dialogues for each phase are presented in Figure 3.

PHASEMI is generated using GPT-4o to instantiate the therapist, client, and supervisor LLMs, and consists of  $N = 13,897$  dialogue instances. The number of samples corresponding to (i) academic, (ii) interpersonal, (iii) emotional, and (iv) career-related concerns is reported in Table 1. Conversation length is measured in therapist–client turns and excludes supervisor turns, although all supervisor interactions are included in the released dataset. Across the dataset, dialogues contain a minimum of 20 turns and a maximum of 44 turns, with an average length of 23.92 turns per dialogue.

	GradSchool	College
Academic	581	2830
Interpersonal	285	2093
Emotional	304	824
Career	1467	5513

Table 1: Category distribution of PHASEMI by data source.

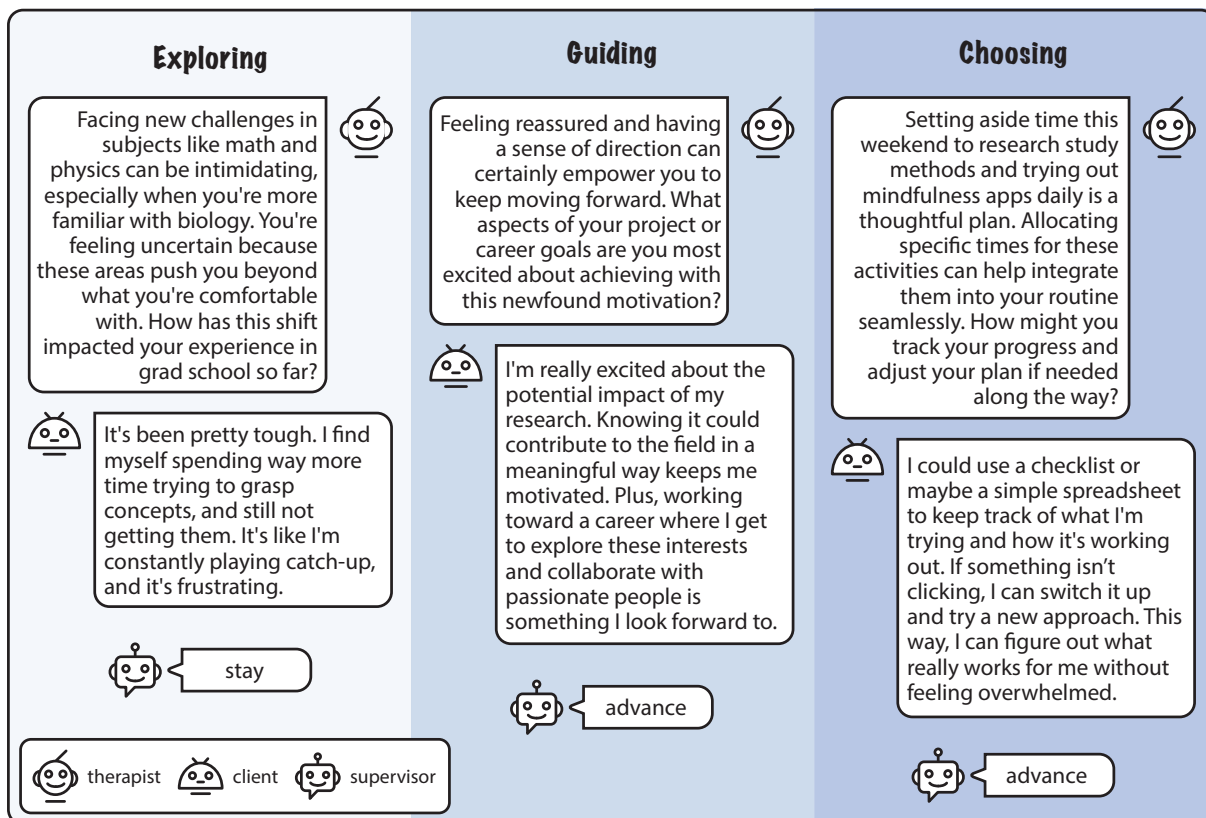


Figure 3: Example single-turn dialogues for each phase. The Exploring phase elicits the client’s experiences and ambivalence, the Guiding phase clarifies motivation and direction, and the Choosing phase emphasizes concrete action and commitment.

## 5 Experiments

We evaluate whether two key criteria are well reflected: (1) counseling quality and (2) phase coverage.

### 5.1 Evaluation Criteria

#### 5.1.1 Counseling Quality

**MITI global rating.** We adopt the Motivational Interviewing Treatment Integrity (MITI<sup>5</sup>) framework as our evaluation protocol, focusing on its global ratings. MITI is an observer-rated coding system designed to assess fidelity to MI and is one of the most widely used evaluation tools in the MI literature (Hurlocker et al., 2020). The global ratings comprise two components: a *technical* component, computed as the average of the cultivating change talk and softening sustain talk scores, and a *relational* component, computed as the average of the partnership and empathy scores. All ratings are measured on an ordinal scale ranging from 1 to 5. Detailed criteria for each component are provided in Section C.1.

<sup>5</sup>We use MITI version 4.2.1.

**General counseling quality.** To assess general counseling quality, we adopt the Cognitive Therapy Rating Scale (CTRS), which is widely regarded as a gold-standard instrument for evaluating both cognitive behavioral therapy (CBT)–specific competencies and general counseling skills. CBT is one of the most commonly used therapeutic approaches in conjunction with MI. Following prior work (Lee et al., 2024), and considering the text-based nature of our dialogues, we focus on three criteria that reflect general counseling skills: *Understanding*, *Interpersonal Effectiveness*, and *Collaboration*. All ratings are measured on an ordinal scale ranging from 0 to 6. Detailed criteria for each component are provided in Section C.2.

#### 5.1.2 MI Phase Coverage

To assess phase coverage for each dialogue, we employ an LLM-as-a-Judge (Claude Sonnet 4) formulated as a binary classification task. We consider three core MI phases grounded in prior work (Resnicow and McMaster, 2012): Exploring, Guiding, and Choosing. Specifically, the judge model is prompted to determine whether the goals

Counseling Quality (Score: 1–5)			
Method	MITI Tech.	MITI Rel.	General
KMI	3.11	2.83	3.19
<b>PHASEMI</b>	<b>3.74</b>	<b>3.81</b>	<b>4.84</b>

MI Phase Coverage (%)			
Method	Exploring	Guiding	Choosing
AnnoMI	87.27	61.82	33.64
KMI	98.00	40.70	11.00
<b>PHASEMI</b>	<b>99.60</b>	<b>99.40</b>	<b>94.70</b>

Table 2: Dataset quality evaluation. The upper section reports average counseling quality scores on a 5-point integer scale (higher is better). MITI Tech. and MITI Rel. denote the MITI Global Rating (Technical and Relational), respectively, while General refers to general counseling quality. The lower section reports MI phase coverage as percentages (%). Results for PHASEMI are computed on the 1,000 sampled subset (not the full set). All values are rounded to two decimal places.

and contextual characteristics of each phase are semantically present in a given dialogue. The classification prompts for the LLM-as-a-Judge are provided in Section F.2.

## 5.2 Experimental Setup

**Baseline.** We use AnnoMI (Wu et al., 2022) and KMI (Kim et al., 2025a) as baseline datasets. AnnoMI comprises 133 dialogues collected from real-world counseling sessions. KMI consists of 1,000 synthetic dialogues generated using a two-model pipeline with LLMs acting as a therapist and a client.

**Counseling quality of dataset.** We evaluate the intrinsic quality of the dataset. Each dialogue is scored on an integer scale using Claude Sonnet 4; the evaluation metrics and prompts are detailed in Section C and F. Client resistancy is a key factor affecting counseling difficulty; however, prior work shows that AI clients—particularly GPT-based models—are overly agreeable, easily persuaded, and insufficiently sensitive to poor counselor performance compared to real-world clients (Cook et al., 2025). Therefore, to ensure fair comparison, we evaluate only KMI and PHASEMI, which employ the same AI client.

**Fine-tuned model quality.** We next evaluate how each dataset contributes to model fine-tuning. We fine-tune Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct on both baseline datasets and our proposed

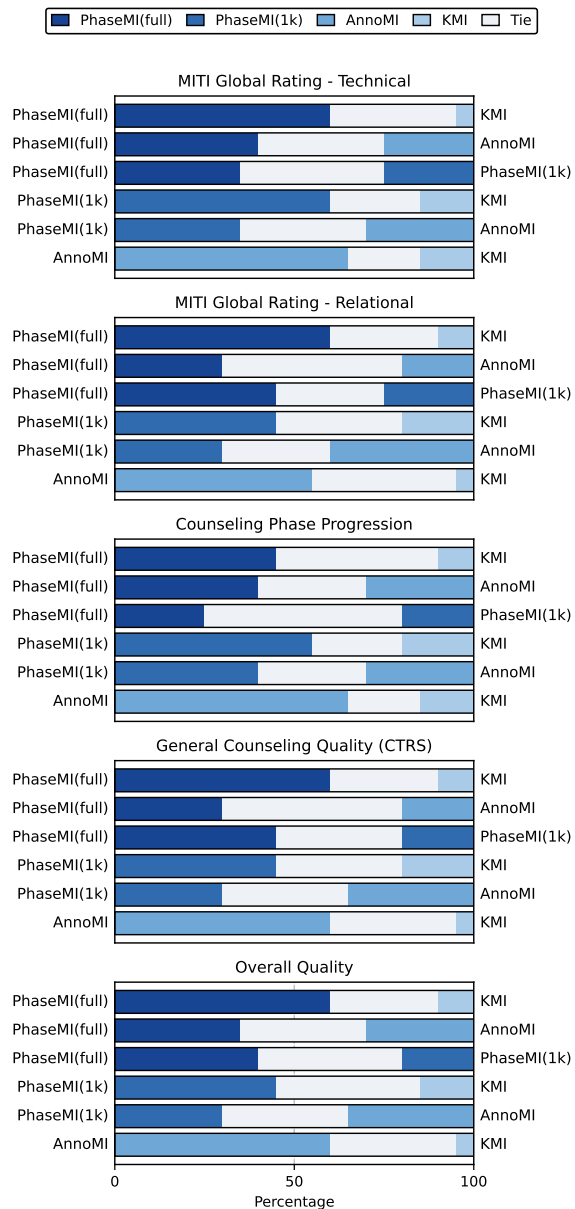


Figure 4: Human expert evaluation. Win rates are computed for AnnoMI, KMI, PHASEMI(1k sampled), and PHASEMI(full). Disagreements between the two human expert raters are treated as ties.

dataset. All models generate test dialogues with the same random seed.

The evaluation set is constructed from two QA-style counseling datasets, MentalChat16k (Xu et al., 2025) and Psych8k (Liu et al., 2023). From MentalChat16k, we retain only samples grounded in real interview data. Client concerns are further filtered for MI suitability using the GPT-4o API. We then randomly sample 30 concerns from each dataset, yielding 60 evaluation cases.

At test time, we simulate real-world usage by terminating dialogue generation once the conversation

<i>Counseling Quality (Score: 1–5)</i>				
<b>Model Size</b>	<b>Dataset</b>	<b>MITI Technical</b>	<b>MITI Relational</b>	<b>General Counseling</b>
7B	AnnoMI	<b>3.76</b>	3.59	4.09
	KMI	3.73	<b>3.61</b>	<b>4.15</b>
	<b>PHASEMI</b>	3.73	<b>3.61</b>	<b>4.15</b>
14B	AnnoMI	3.92	3.72	<b>4.61</b>
	KMI	3.94	<b>3.81</b>	4.55
	<b>PHASEMI</b>	<b>3.98</b>	3.79	<b>4.61</b>

<i>MI Phase Coverage (%)</i>				
<b>Model Size</b>	<b>Dataset</b>	<b>Exploring</b>	<b>Guiding</b>	<b>Choosing</b>
7B	AnnoMI	96.67	70.00	50.00
	KMI	95.00	63.33	58.33
	<b>PHASEMI</b>	<b>100.00</b>	<b>71.67</b>	<b>68.33</b>
14B	AnnoMI	<b>98.33</b>	73.33	50.00
	KMI	<b>98.33</b>	78.33	51.67
	<b>PHASEMI</b>	<b>98.33</b>	<b>83.33</b>	<b>60.00</b>

Table 3: Evaluation on test dialogues generated by fine-tuned models. The upper section reports average LLM-as-a-Judge scores for counseling quality. MITI Technical and MITI Relational denote the MITI Global Rating (Technical and Relational), respectively, while General Counseling refers to general counseling quality. The lower section reports MI phase coverage as percentages (%). For fairness, PHASEMI refers to a model trained on a 1,000-sampled subset to match dataset size of KMI. All values are rounded to two decimal places.

naturally concludes. A terminator module detects appropriate stopping points, and no minimum turn constraints are imposed. Both LLM-as-a-Judge (Claude Sonnet 4) and human expert evaluations<sup>6</sup> are conducted.

**Model training details.** Following the baseline configuration of Kim et al. (2025a), models trained on AnnoMI and KMI are fine-tuned for 3 epochs. We also train the PHASEMI 1k-sampled variant, which contains 1,000 instances to match the scale of KMI, for 3 epochs under the same setting. In contrast, the full PHASEMI set, which contains 13,897 instances, is trained for 1 epoch due to its substantially larger size. We use batch training with gradient accumulation, apply LoRA for parameter-efficient fine-tuning, and train the models in bfloat16 precision. Detailed training configurations are provided in Table 5.

## 5.3 Results

### 5.3.1 Dataset Quality

For all the criteria regarding counseling quality, PHASEMI significantly outperforms KMI (Table 2 Top). For MI phase coverage, PHASEMI covers all

<sup>6</sup>Evaluation is conducted on a subset of 20 samples from the 60-sample test set due to annotation constraints. It is performed by two licensed clinical psychologists (Ph.D.) with extensive clinical experience.

phases, whereas the baseline datasets primarily focus on the Exploring phase and rarely transition to later phases (Table 2 Bottom). In particular, baseline datasets seldom reach the Choosing phase, indicating that counseling sessions are rarely brought to a proper conclusion.

### 5.3.2 Fine-tuned Model Quality

PHASEMI demonstrates improved performance across all evaluation criteria under more stringent assessments by certified professional human experts (Figure 4), while maintaining comparable counseling quality to models trained on other baseline datasets in the LLM-as-a-Judge evaluation (Table 3). The evaluation guideline for human experts is provided in Section J.

For MI phase coverage, while models fine-tuned on any datasets adequately exhibit the Exploring phase, those trained on baseline datasets show limited coverage of the Guiding and Choosing phases (Table 3). Although the model fine-tuned on PHASEMI does not reach the Choosing phase in every case, it demonstrates substantially better phase coverage than the baselines.

## 6 Closing Remark

In this work, we present PHASEMI, a phase-structured MI dataset designed to support phase-

aware LLM-based counseling systems. We introduce a data generation framework that explicitly controls MI phase transitions through the interaction among therapist, client, and supervisor LLMs, enabling coherent and theory-aligned dialogue progression. Our results show that PHASEMI provides more structured supervision than existing MI datasets in terms of both counseling quality and MI phase coverage. Although PHASEMI is synthetic and does not aim to validate deployment in real counseling settings, this controlled design is necessary to study phase-level supervision, which is difficult to isolate from real-world transcripts with implicit and noisy phase boundaries.

PHASEMI also suggests a practical path toward real-world counseling support. At the dataset level, it provides structured supervision for training models to better capture counseling progression. At the model level, PHASEMI-trained models may support counselor training by providing feedback on phase progression and overall session flow, beyond local technique-level feedback alone (Yang et al., 2024; Louie et al., 2026). We hope PHASEMI serves as a foundation for future research on phase-aware and clinically grounded counseling support systems.

## Limitations

Client resistance is a key factor influencing the difficulty of counseling progression. Prior work suggests that AI clients, particularly GPT-based models, often remain overly agreeable, are easily persuaded by counselor suggestions, and seldom exhibit confusion or resistance even under poor therapist performance (Cook et al., 2025). Accordingly, our framework should be understood as a theory-guided pipeline for structuring counseling progression, rather than as an attempt to fully reproduce the behavioral complexity of real-world counseling interactions. Our design therefore prioritizes the faithful operationalization of MI theory over exhaustive modeling of client-side behavioral variability. Incorporating client resistance remains an important direction for future work.

## Acknowledgments

We thank Professor Ho-Jin Choi, Gyouk Chu, Hangyul Yoon, Minyeong Hwang, Sihwan Park and Joonhyung Park for their valuable suggestions and discussions, which helped improve the clarity of the manuscript. We gratefully acknowledge

the support from the InnoCORE Program of the Ministry of Science and ICT (No. N10250156).

## References

- Anzar Abbas, Vijay Yadav, Emma Smith, Elizabeth Ramjas, Sarah B Rutter, Caridad Benavidez, Vidya Koesmahargyo, Li Zhang, Lei Guan, Paul Rosenfield, and 1 others. 2021. Computer vision-based assessment of motor functioning in schizophrenia: Use of smartphones for remote measurement of schizophrenia symptomatology. *Digital Biomarkers*, 5(1):29–36.
- Nicholas W Affrunti and Torrey A Creed. 2019. The factor structure of the cognitive therapy rating scale (ctrs) in a sample of community mental health clinicians. *Cognitive Therapy and Research*, 43(3):642–655.
- Ankit Aich, Avery Quynh, Varsha Badal, Amy Pinkham, Philip Harvey, Colin Depp, and Natalie Parde. 2022. Towards intelligent clinically-informed language analyses of people with bipolar disorder and schizophrenia. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2871–2887.
- Eliane M Boucher, Nicole R Harake, Haley E Ward, Sarah Elizabeth Stoeckl, Junielly Vargas, Jared Minkel, Acacia C Parks, and Ran Zilca. 2021. Artificially intelligent chatbots in digital mental health interventions: a review. *Expert review of medical devices*, 18(sup1):37–49.
- Keng-hao Chang, Matthew K Chan, and John Canny. 2011. Analyzethis: Unobtrusive mental health monitoring by voice. In *CHI’11 Extended Abstracts on Human Factors in Computing Systems*, pages 1951–1956.
- David A Cook, Joshua Overgaard, V Shane Pankratz, Guilherme Del Fiol, and Chris A Aakre. 2025. Virtual patients using large language models: Scalable, contextualized simulation of clinician-patient dialogue with feedback. *Journal of Medical Internet Research*, 27:e68486.
- Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.
- John Gideon, Emily Mower Provost, and Melvin McInnis. 2016. Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 2359–2363. IEEE.
- Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters,

- Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.
- Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, and 1 others. 2025. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Margo C Hurlocker, Michael B Madson, and Julie A Schumacher. 2020. Motivational interviewing quality assurance: A systematic review of assessment tools across research contexts. *Clinical psychology review*, 82:101909.
- Melanie M Iarussi. 2013. Examining how motivational interviewing may foster college student development. *Journal of College Counseling*, 16(2):158–175.
- Sharon James, Susan Mcinnes, Elizabeth Halcomb, and Jane Desborough. 2020. General practice nurses’ communication strategies for lifestyle risk reduction: A content analysis. *Journal of advanced nursing*, 76(11):3082–3091.
- Zifan Jiang, Sahar Harati, Andrea Crowell, Helen S Mayberg, Shamim Nemati, and Gari D Clifford. 2020. Classifying major depressive disorder and response to deep brain stimulation over time by analyzing facial expressions. *IEEE transactions on biomedical engineering*, 68(2):664–672.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.
- Hyunjong Kim, Suyeon Lee, Yeongjae Cho, Eunseo Ryu, Yohan Jo, Suran Seong, and Sungzoon Cho. 2025a. **KMI: A dataset of Korean motivational interviewing dialogues for psychotherapy**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10803–10828, Albuquerque, New Mexico. Association for Computational Linguistics.
- Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, and 1 others. 2025b. Pubchem 2025 update. *Nucleic acids research*, 53(D1):D1516–D1525.
- Suyeon Lee, Sunghwan Mac Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, and 1 others. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274.
- June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.
- Ryan Louie, Raj Sanjay Shah, Ifdita Hasan Orney, Juan Pablo Pacheco, Emma Brunskill, and Diyi Yang. 2026. Can llm-simulated practice and feedback up-skill human counselors? a randomized study with 90+ novice counselors. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–31.
- Brad Lundahl and Brian L Burke. 2009. The effectiveness and applicability of motivational interviewing: A practice-friendly review of four meta-analyses. *Journal of clinical psychology*, 65(11):1232–1245.
- William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.
- William R Miller and Gary S Rose. 2009. Toward a theory of motivational interviewing. *American psychologist*, 64(6):527.
- Saif Mohammad. 2024. Worrywords: Norms of anxiety association for over 44k english words. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16261–16278.
- Paola Pedrelli, Maren Nyer, Albert Yeung, Courtney Zulauf, and Timothy Wilens. 2015. College students: mental health problems and treatment considerations. *Academic psychiatry*, 39(5):503–511.
- Jinghui Qin, Changsong Liu, Tianchi Tang, Dahuang Liu, Minghao Wang, Qianying Huang, and Rumin Zhang. 2025. Mental-perceiver: Audio-textual multimodal learning for estimating mental disorders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25029–25037.
- Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636.

- Ken Resnicow and Fiona McMaster. 2012. Motivational interviewing: moving from why to how with autonomy support. *International Journal of Behavioral Nutrition and Physical Activity*, 9(1):19.
- Gary R VandenBos. 2007. *APA dictionary of psychology*. American Psychological Association.
- Lei Wang, Kathryn I Alpert, Vince D Calhoun, Derin J Cobia, David B Keator, Margaret D King, Alexandr Kogan, Drew Landis, Marcelo Tallis, Matthew D Turner, and 1 others. 2016. Schizconnect: Mediating neuroimaging databases on schizophrenia and related disorders for large-scale integration. *Neuroimage*, 124:1155–1167.
- Lei Wang, Alex Kogan, Derin Cobia, Kathryn Alpert, Anthony Kolasny, Michael I Miller, and Daniel Marcus. 2013. Northwestern university schizophrenia data and software tool (nusdast). *Frontiers in neuroinformatics*, 7:25.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambadur, David Rosenberg, and Gideon Mann. 2023a. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Yuqi Wu, Jie Chen, Kaining Mao, and Yanbo Zhang. 2023b. Automatic post-traumatic stress disorder diagnosis via clinical transcripts: a novel text augmentation with large language models. In *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–5. IEEE.
- Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. 2023c. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3):110.
- Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181. IEEE.
- Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5367–5378.
- Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*.
- Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *FinLLM at IJCAI*.
- JE Young and Aaron T Beck. 1980. Cognitive therapy scale. *Unpublished manuscript, University of Pennsylvania*.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. In *ICLR*.
- Zhouan Zhu, Chenguang Li, Jicai Pan, Xin Li, Yufei Xiao, Yanan Chang, Feiyi Zheng, and Shangfei Wang. 2023. Medic: A multimodal empathy dataset in counseling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6054–6062.

## A Counseling Terminology

This section introduces key counseling-related terminology used throughout the paper. The definitions are primarily grounded in established literature on MI and counseling psychology, drawing in particular from Miller and Rollnick (2012) and VandenBos (2007).

**Change talk.** Any client speech that favors movement toward a specific change goal or target behavior. Change talk may refer to both early consideration of change (*e.g.*, expressing desire or intention) and language that reflects commitment or steps toward actual behavior change.

**Sustain talk.** Any client speech that favors maintaining the status quo rather than moving toward change. Sustain talk is sometimes referred to as *counterchange talk* and is always defined with respect to a specific change target. Such utterances typically emphasize comfort, perceived benefits, or necessity of the current behavior.

**Status Quo.** The current state of affairs in which no change has been initiated or pursued.

Talk type	Example Statement
Change Talk	“I’m willing to give it another try.”
Sustain Talk	“It’s the only way I have to relax.”

Table 4: Illustrative examples of change talk and sustain talk in the context of tobacco cessation.

**Ambivalence.** The simultaneous presence of competing motivations both for and against change. Ambivalence is also conceptualized as an internal debate between arguments supporting change and arguments favoring the status quo. Importantly, ambivalence is considered a normal part of the change process rather than a form of resistance or pathology. In practice, change talk and sustain talk may co-occur within a single utterance, *e.g.*, “I wish I could quit smoking, but I don’t think I can.”

**Presenting problem.** The problem or symptom identified by the client as the primary reason for seeking treatment. In psychotherapy, presenting problems may include depression, anxiety, panic, anger, chronic pain, or interpersonal difficulties. Such problems may become the focus of treatment or may reflect underlying issues not initially recognized by the client.

**Evidence-based treatment (EBT).** A therapeutic method for which there is sufficient empirical evidence demonstrating efficacy, based on established scientific and methodological criteria.

**Resistance.** Traditionally defined as unconscious defensive processes that inhibit the emergence of threatening material. In some clinical frameworks, resistance has also been used to describe or attribute clients’ noncompliance or oppositional responses.

## B Presenting Problem Taxonomy

This section outlines the taxonomy of presenting problems used to filter and categorize the seed concerns in our dataset. The taxonomy is adapted from prior counseling literature (Miller and Rollnick, 2012; Iarussi, 2013).

**Academic Concerns.** Academic concerns primarily relate to academic competence and time management. This category includes anxiety regarding overall performance, chronic procrastination or lateness, habitual non-completion of required tasks, and inadequate preparation for expected activities.

**Interpersonal Concerns.** Interpersonal concerns involve difficulties in forming, maintaining, or exiting relationships. Representative cases include ambivalence about leaving violent or toxic relationships, boundary-setting conflicts, social withdrawal, mistrust toward peers, and persistent feelings of loneliness.

**Emotional Concerns.** Emotional concerns encompass challenges in emotion regulation and identity formation. Issues related to emotion management include difficulty accepting or regulating negative emotions, recurrent anger outbursts with potential disciplinary consequences, and substance use as an avoidance strategy. This category also covers identity-related concerns such as distress regarding body image, gender identity, or sexual orientation, uncertainty about personal roles, lifestyle, or cultural identity, and low self-acceptance or self-esteem.

**Career Concerns.** Career concerns reflect uncertainty in both short-term career decisions and long-term career trajectories. This category includes weak commitment to activities or interests and resulting misalignment between goals and behaviors.

## C Counseling Quality Evaluation Metrics

### C.1 Motivational Interviewing Treatment Integrity (MITI)

MITI is the most widely adopted behavioral coding system for evaluating a counselor's adherence to the principles and techniques of MI. MITI comprises two components: global scores and behavior counts. In this paper, we exclusively focus on the global scores, as fine-grained behavior counting lies beyond the scope of our work.

Global scores assign a single rating on a five-point Likert scale to capture the overall quality of an interaction between a counselor and a client. These scores reflect the evaluator's holistic assessment across four dimensions: Cultivating Change Talk, Softening Sustain Talk, Partnership, and Empathy. These dimensions are further organized into two higher-level categories: Technical Global and Relational Global. Following standard practice, the Technical Global score is computed as the average of Cultivating Change Talk and Softening Sustain Talk, while the Relational Global score is obtained by averaging Partnership and Empathy (Miller and Rollnick, 2012).

#### C.1.1 Technical Global

**Cultivating Change Talk.** This dimension assesses the extent to which the counselor actively elicits and reinforces the client's own language in favor of the change goal, including expressions of motivation, commitment, and confidence for change. Counselors who demonstrate a consistent and intentional effort to increase the depth, strength, or momentum of client change talk receive higher ratings.

**Softening Sustain Talk.** This dimension evaluates the extent to which the counselor avoids emphasizing reasons for maintaining the status quo or resisting change. Higher ratings are assigned to counselors who consistently work to reduce the depth, strength, or momentum of the client's sustain talk, thereby minimizing reinforcement of arguments against change.

#### C.1.2 Relational Global

**Partnership.** This dimension measures the degree to which the counselor communicates an understanding that expertise and wisdom regarding change primarily reside within the client. Counselors who actively foster collaboration and power

sharing—such that the client's contributions meaningfully shape the direction and content of the session—receive higher ratings.

**Empathy.** This dimension assesses the counselor's ability to understand, or make a genuine effort to understand, the client's perspective and lived experience. Higher ratings reflect evidence of deep empathic understanding, including sensitivity not only to what the client explicitly states but also to underlying meanings and unspoken experiences.

### C.2 Cognitive Therapy Rating Scale (CTRS)

CBT is one of the most extensively studied forms of psychotherapy, alongside Motivational Interviewing. CTRS is among the most widely adopted instruments for assessing treatment fidelity in CBT (Affrunti and Creed, 2019). The CTRS was originally developed to capture two theorized dimensions of therapist competence: general therapeutic skills and cognitive-behavioral techniques (Young and Beck, 1980).

In this paper, we focus on a subset of CTRS dimensions that reflect general counseling competencies as mentioned above—*Understanding*, *Interpersonal Effectiveness*, and *Collaboration*—which capture the quality of the therapeutic relationship and the counselor's general interpersonal skills.

**Understanding.** This dimension evaluates the extent to which the counselor is able to empathically enter the client's experiential world, accurately perceive the client's perspective, and communicate this understanding effectively. High ratings reflect sensitivity not only to the client's explicit verbal expressions but also to implicit cues conveyed through affect, tone, and non-verbal behavior, recognizing that clients may be unable or unwilling to fully articulate their internal experiences.

**Interpersonal Effectiveness.** This dimension assesses whether the counselor demonstrates appropriate levels of warmth, concern, confidence, genuineness, and professionalism. Effective interpersonal engagement is reflected in the counselor's ability to appear sincere, authentic, and open, thereby fostering a supportive and credible therapeutic presence.

**Collaboration.** This dimension measures the degree to which the counselor and client engage in a cooperative and goal-aligned working relationship. Collaboration is often manifested as a therapeutic alliance in which both parties jointly address the

Hyperparameter	Value
LoRA Rank ( $r$ )	64
LoRA Alpha	32
LoRA Dropout	0.05
Target Modules	{q_proj, v_proj}
Learning Rate	1e-6
Precision	bf16
Assistant-only Loss	True

Table 5: Hyperparameters for LoRA-based supervised fine-tuning with TRL SFTTrainer.

client’s difficulties. A strong collaborative stance helps ensure alignment of goals throughout the session, reduces client resistance, and minimizes the risk of misunderstandings between the counselor and the client.

## D Prompt Templates for Data Generation

This section summarizes the prompt templates employed for dialogue generation in both the main dataset construction and the test dialogue generation used for model evaluation.

### D.1 Dataset Generation

Tables 7, 8, 9, and 10 present the prompt templates and shared variables used in the dataset generation process.

### D.2 Test Dialogue Generation

Table 11 presents the prompt template used to generate the test dialogues for model evaluation.

## E Training Details

Table 5 summarizes the hyperparameters used for LoRA fine-tuning. We use the SFTTrainer from TRL for supervised fine-tuning and compute the loss only on assistant responses.

## F Prompt Templates for Evaluation

This section summarizes the prompt templates employed for accessing counseling quality of data and the phase coverage of the data.

### F.1 Counseling Quality Evaluation

Tables 13, 14, and 15 present the prompt templates employed to access counseling quality.

### F.2 Phase Coverage Classification

Table 16 presents the prompt template employed to evaluate MI phase coverage.

## G Cost Analysis on Data Generation

We analyze the cost of data generation and evaluation using the OpenAI gpt-4o API. For dataset construction, generating 1,000 dialogues with batch processing required less than \$50.

For evaluation, test dialogues were generated using a real-time API setting, where 60 dialogues were sampled per model. The cost was approximately \$2 per model for generating these 60 dialogues, where the OpenAI gpt-4o API was used alongside the corresponding fine-tuned models.

Overall, the results indicate that both dataset construction and evaluation can be conducted at relatively low cost, making the proposed pipeline practical for iterative experimentation.

## H Example Dialogues from PHASEMI

This section further provides example dialogues from PHASEMI, as shown in Tables 17 and 18.

## I Example Test Dialogues for Model Quality Evaluation

This section presents example test dialogue triplets generated from a shared client concern seed using different fine-tuned models. Specifically, the dialogues are produced by models trained on AnnoMI (Tables 19 and 20), KMI (Table 21), and PHASEMI (Tables 22 and 23). The client concern seed used to generate these dialogues is shown in Table 6.

I know you’re just trying to help, but every time I talk to someone like you, they always tell me the same thing – I need to get my grades up, stop smoking pot, go back to class, and do work. It’s frustrating because I’ve heard it so many times, and it feels like nobody understands.

Table 6: Client’s concern seed utilized to generate the following examples described in Tables 19, 20, 21, 22, and 23.

## J Instructions Given to Human Evaluators

Tables 24, 25, and 26 present the instruction given to human evaluators.

## **K The Use of Large Language Models (LLMs)**

We used ChatGPT (GPT-5, OpenAI) exclusively to aid with writing and polishing the text, such as improving grammar, fluency, and clarity of exposition. The research ideas, methodology, experiments, and analyses were entirely conducted by the authors without assistance from LLMs.

<b>Phase</b>	<b>Therapist Prompt Template</b>
<b>Exploring</b>	<p>You are an MI counselor in the Exploring phase. [MI_phases_overview]</p> <p>Comfort the client to build rapport and elicit the client’s story, including their feelings and relevant context. Collaboratively identify which behaviors to address. Explore behavioral change history if available.</p> <p>Use open-ended questions and reflections:</p> <p>Content reflection: paraphrase without parroting.</p> <p>Feeling reflection: explain why they feel a certain way (“You feel X because of Y”).</p> <p>Double-sided reflection: acknowledge both sides of ambivalence without pushing change.</p> <p>Stay in the Exploring phase; avoid cultivating change talk. [Tone]</p> <p>Reply in 1–2 short, natural sentences. Continue the given conversation turn.</p>
<b>Guiding</b>	<p>You are an MI counselor in the Guiding phase. [MI_phases_overview]</p> <p>Encourage change talk by exploring life with and without change and highlighting gaps between current actions and life goals.</p> <p>Use 0–10 confidence rulers, values clarification, and summarizing.</p> <p>Stay in Guiding; avoid detailed action planning.</p> <p>Avoid excessive questions. [Tone]</p> <p>Reply in 1-2 short, natural sentences, continuing the given conversation turn.</p>
<b>Choosing</b>	<p>You are an MI counselor in the Choosing phase. [MI_phases_overview]</p> <p>Help the client set a goal, create an action plan, anticipate barriers, and agree on monitoring.</p> <p>Use menu building, goal setting, and action reflections based on steps the client mentioned or implied, offering options to support autonomy.</p> <p>e.g., “Based on what you said, possible options include X and Y.” [Tone]</p> <p>Reply in 1-2 short, natural sentences, continuing the given conversation turn.</p>
<b>Terminating</b>	<p>You are an MI counselor closing the counseling session.</p> <p>Your mission is to summarize key goals, action steps, and coping strategies.</p> <p>Offer encouragement and a clear next step support plan.</p> <p>Avoid questions and respond in 1-2 short, natural and conversational style. [Tone]</p> <p>Continue the given conversation turn.</p>

Table 7: Therapist prompt template used for data generation. Shared variables (MI\_phases\_overview and Tone) are defined in Table 10.

<b>Phase</b>	<b>Client Prompt Template</b>
<b>Exploring</b>	You are a U.S. student in university who posted this concern on Reddit: [concern].
<b>Guiding</b>	You’re in counseling with a school therapist.
<b>Choosing</b>	Respond to their questions in 2 sentences, reflecting your concern without repeating them or giving details. Continue the conversation as the client, speaking in a casual, unstructured tone.
<b>Terminating</b>	You are a U.S. student in university about to finish a counseling session with the school therapist. Respond to their suggestions in 1–2 simple sentences, reflecting your concern without repeating them or giving details. Continue the conversation as the client, in a casual, unstructured tone.

Table 8: Client prompt template used for data generation.

Phase	Supervisor Prompt Template
<b>Exploring</b>	<p>You are an MI progress evaluator. [MI_phases_overview]  CURRENT_PHASE: Exploring  Decide if ready to advance to Guiding.  INPUTS: [Conversation history] [Client's concern]  RULES: Exploring goals met: essence of client's story elicited including the client's thought, detailed relevant context behind the client's thought and feelings, empathy &amp; rapport, mutually agreed goals appeared. [Turns_recommendation]  Decision: Determine conservatively in favor of staying in Exploring.  [Decision_detail] [Output_format]</p>
<b>Guiding</b>	<p>You are an MI progress evaluator. [MI_phases_overview]  CURRENT_PHASE: Guiding  Decide if ready to advance to Choosing.  INPUTS: [Conversation history] [Client's concern]  RULES: Guiding goals met:</p> <ul style="list-style-type: none"> <li>• Discrepancy between behavior and values/goals evoked</li> <li>• Explored life with and without change</li> <li>• Used confidence rulers or values clarification</li> <li>• Summaries given and client reacted</li> </ul> <p>Readiness signals:</p> <ul style="list-style-type: none"> <li>• Change Talk present</li> <li>• At least one specific change goal or next step</li> <li>• Confidence <math>\geq 6</math> (if rulers used)</li> <li>• Ambivalence or sustain talk decreased [Turns_recommendation]</li> </ul> <p>Decision: Determine conservatively in favor of staying in Guiding.  [Decision_detail] [Output_format]</p>
<b>Choosing</b>	<p>You are an MI progress evaluator. [MI_phases_overview]  CURRENT_PHASE: Choosing  Decide if ready to advance to Terminating.  INPUTS: [Conversation history] [Client's concern]  RULES: Choosing goals met:</p> <ul style="list-style-type: none"> <li>• Clear goal and concrete action plan</li> <li>• Barriers discussed and self-monitoring agreed</li> </ul> <p>Readiness signals:</p> <ul style="list-style-type: none"> <li>• Commitment and taking-steps talk</li> <li>• Action reflections accepted or refined</li> <li>• Confidence in ability to change [Turns_recommendation]</li> </ul> <p>Decision: Determine conservatively in favor of staying in Choosing.  [Decision_detail] [Output_format]</p>
<b>Terminating</b>	<p>You are an MI progress evaluator closing the session. Decide whether to conclude the entire counseling session ("stop") or remain in Termination ("continue").  GOAL: Decide to end the session ("stop") or remain in Termination ("continue").  INPUTS: [Conversation history] [Client's concern]  RULES: Termination goals met:</p> <ul style="list-style-type: none"> <li>• Summary of goals, steps, and coping strategies</li> <li>• Next-step plan discussed</li> </ul> <p>Decision:</p> <ul style="list-style-type: none"> <li>• If natural to end or stuck in circles <math>\rightarrow</math> "move"</li> <li>• If no closure invitation <math>\rightarrow</math> "continue"</li> <li>• A lenient evaluation is recommended</li> </ul> <p>[Decision_detail] [Output_format]</p>

Table 9: Supervisor prompt templates used for data generation. The variables (Conversation history, Client's concern) are provided as runtime user inputs. All shared prompt variables (MI\_phases\_overview, Turns\_recommendation, Decision\_detail, and Output\_format) are defined in Table 10.

Variable	Content
MI_phases _overview	MI Phases Overview 1. Exploring: Initial counseling phase that elicits the client’s story and builds rapport. 2. Guiding: Strengthen motivation for change and evoke Change Talk. 3. Choosing: Collaboratively develop specific action plans and strategies for anticipated barriers.
Tone	Use a warm, empathetic, non-judgemental, encouraging tone throughout the conversation. Use OARS (Open-ended questions, Affirmations, Reflective listening, Summarizing) as primary communication techniques. Include these techniques in every response. Establish a non-confrontational climate in which clients can feel comfortable expressing their thoughts. Do not make direct attempt to dismantle denial or confront irrational, and rarely attempt to persuade. Encourage clients to make fully informed and deeply contemplated decisions. Instead of always starting with ‘It sounds like’ or ‘It seems like,’ make use of diverse expressions.
Turns_ recommendation	Each phase should include at least 3-4 turns per speaker to effectively achieve the goals.
Decision_ detail	<ul style="list-style-type: none"> <li>• If met: move</li> <li>• If stuck in repetitive, unproductive questions: move</li> <li>• Else: continue</li> </ul>
Output_ format	OUTPUT JSON only: { { “reasoning”: “<max 80 words explanation, but not mentioning the number of turns>”, “determination”: “move”   “continue” } }

Table 10: Shared prompt variables referenced in Table 7 and Table 9.

Model	Prompt Template
<b>Therapist</b>	You are a counselor specialized in Motivational Interviewing (MI). You will be provided with a conversation between a therapist and a client. Generate the therapist’s next utterance, ensuring it aligns with the principles of Motivational Interviewing. Do not simulate the client’s response, only the therapist’s next utterance. [Conversation history]
<b>Client</b>	You posted this concern: [concern]. You’re in counseling with a therapist. Respond to their questions in 2 sentences, reflecting your concern without repeating them or giving details. Continue the conversation as the client, speaking in a casual, unstructured tone. Do not simulate the therapist’s response, only the client’s next utterance. [Conversation history]
<b>Terminator</b>	You are a counselor specialized in Motivational Interviewing (MI). Your task is to judge whether the session has reached a natural and appropriate ending point. Focus on the situation, not the dialogue content itself. A natural ending typically includes one of the following: Summarizing or closing the discussion, Setting a future appointment or planning a follow-up. Respond strictly in JSON format as follows: { “reason”: “A concise explanation (within 50 words) of why you made this judgment.”, “determination”: “continue   terminate” } [Conversation history]

Table 11: Prompt templates used to generate test dialogues. The [Conversation history] is provided as a runtime user input, while the [concern] is instantiated from seed data.

---

You are a professional counselor.  
 You will be provided with a transcript of a counseling session.  
 Your task is to assess the therapist on a scale from {min} to {max} based on the given criteria.  
 Please follow these strict instructions:

1. Read the counseling session transcript carefully.
2. Review the evaluation question and criteria thoroughly.
3. Grade conservatively based on the criteria.
  - If there is any deficiency, no matter how small, assign a score of 4 or lower.
4. Output only a JSON object in the following format:  

```
{{"score": <INT>, "reason": "<concise explanation>"}}
```

Example output:  

```
{{"score": 3, "reason": "Therapist did not sufficiently elicit change talk and lacked reflective listening."}}
```

Do not include any additional text. No prefixes, headers, or extra whitespace.

Evaluation description [description]  
 Rating scale [rating\_scale]

---

Table 12: Prompt Template for Counseling Quality evaluation. min and max refers to (1, 5) for MITI and (0, 6) for General Counseling Quality (or CTRS). description and rating\_scale are provided in the following Table.

Criteria	Content	
Cultivating Change Talk	description	Cultivating change talk measures the extent to which the clinician actively encourages the client’s own language in favor of the change goal, and confidence for making that change. To achieve higher ratings on the Cultivating Change Talk scale, the change goal must be obvious in the session and the conversation must be largely focused on change, with the therapist actively cultivating change talk when possible. Low scores on this scale occur when the therapist is inattentive to the client’s language about change, either by failing to recognize and follow up on it, or by prioritizing other aspects of the interaction (such as history-taking, assessment or nondirective listening).
	rating scale	Assume a default score 3 and moves up or down as indicated. 2 or less: when the client shows inconsistent growth in motivation and continues to hesitate or resist change. 3: when the therapist often attends to the client’s language in favor of change, but misses some changes to encourage change talk. 4 or more: only when the client clearly moves from low or ambivalent motivation at the start to a strong, sustained commitment to change by the end—an unmistakable, dramatic increase.
Softening Sustain Talk	description	Softening sustain talk measures the extent that the therapist avoids a focus on the reasons against changing or for maintaining the status quo. To achieve high scores, therapists should avoid lingering in discussions concerning the difficulty or undesirability of change. Therapists low on the scale focuses considerable attention to the barriers of change, even when using MI-consistent techniques (e.g., asking open questions, offers reflections, affirmations and other MI Adherent techniques) to evoke and reflect sustain talk throughout the session.
	rating scale	Assume a default score 3 and moves up or down as indicated. 2 or less: when the therapist usually chooses to explore, focus on, or respond to the client’s language in favor of the status quo. 3: when the therapist 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 or more: when the therapist shows a marked and consistent effort to decrease the depth, strength, or momentum of the clients language in favor of the status quo.
Partnership	description	Partnership measures the extent to which the clinician conveys an understanding that expertise and wisdom about change reside mostly within the client. Therapists high on this scale behave as if the interview is occurring between two equal partners, both of whom have knowledge that might be useful in solving the change under consideration. Therapists low on the scale assume the expert role for a majority of the interaction and have a high degree of influence in the nature of the interaction.
	rating scale	Assume a default score 3 and moves up or down as indicated. 2 or less: when the therapist superficially responds to opportunities to collaborate. 3: when the therapist incorporates client’s contributions but does so in a lukewarm or erratic fashion. 4 or more: when the therapist actively fosters and encourages power sharing in the interaction in such a way that client’s contributions substantially influence the nature of the session.

Table 13: Prompt used to evaluate counseling quality based on the Motivational Interviewing Treatment Integrity (MITI). The remainder of the table is presented in Table 14.

Criteria	Content
Empathy	<p>description    Empathy measures the extent to which the clinician understands or makes an effort to grasp the client’s perspective and experience (i.e., how much the clinician attempts to “try on” what the client feels or thinks). Reflective listening is an important part of this characteristic, but this global rating is intended to capture all efforts that the clinician makes to understand the client’s perspective and convey that understanding to the client. Therapists high on this scale show evidence of understanding the client’s worldview in a variety of ways including complex reflections that seem to anticipate what clients mean but have not said, insightful questions based on previous listening and accurate appreciation for the client’s emotional state. Therapists low on the scale do not appear interested in the client’s viewpoint.</p> <p>rating scale    Assume a default score 3 and moves up or down as indicated. 2 or less: when the therapist makes sporadic efforts to explore the client’s perspective. Also, when the therapist’s understanding is inaccurate or detracts from the client’s true meaning, this should be reflected in the rating. 3: when the therapist actively tries to understand the client’s perspective, with modest success. 4 or more: when the therapist shows evidence of deep understanding of the client’s perspective, not just for what has been explicitly stated but what the client means but has not yet said.</p>

Table 14: Prompt used to evaluate counseling quality based on the Motivational Interviewing Treatment Integrity (MITI). This table continues from Table 13.

<b>Criteria</b>	<b>Content</b>
Understanding	<p>description      How accurately does the therapist demonstrate understanding of the client’s issues and concerns?</p> <p>rating scale      Descriptions are provided for even numbered scale points. If the therapist falls between two of the descriptors, select the intervening odd number. 0: Therapist repeatedly failed to understand what the patient explicitly said and thus consistently missed the point. Poor empathic skills. 2: Therapist was usually able to reflect or rephrase what the patient explicitly said, but repeatedly failed to respond to more subtle communication. Limited ability to listen and empathize. 4: Therapist generally seemed to grasp the patient’s “internal reality” as reflected by both what the patient explicitly said and what the patient communicated in more subtle ways. Good ability to listen and empathize. 6: Therapist seemed to understand the patient’s “internal reality” thoroughly and was adept at communicating this understanding through appropriate verbal and non-verbal responses to the patient (e.g., the tone of the therapist’s response conveyed a sympathetic understanding of the client’s “message”). Excellent listening and empathic skills</p>
Interpersonal effectiveness	<p>description      How effective is the therapist in maintaining a positive and therapeutic relationship with a client?</p> <p>rating scale      Descriptions are provided for even numbered scale points. If the therapist falls between two of the descriptors, select the intervening odd number. 0: Therapist had poor interpersonal skills. Seemed hostile, demeaning, or in some other way destructive to the patient. 2: Therapist did not seem destructive, but had significant interpersonal problems. At times, therapist appeared unnecessarily impatient, aloof, insincere or had difficulty conveying confidence and competence. 4: Therapist displayed a satisfactory degree of warmth, concern, confidence, genuineness, and professionalism. No significant interpersonal problems. 6: Therapist displayed optimal levels of warmth, concern, confidence, genuineness, and professionalism, appropriate for this particular patient in this session.</p>
Collaboration	<p>description      To what extent does the therapist engage the client in collaborative goalsetting and decision-making?</p> <p>rating scale      Descriptions are provided for even numbered scale points. If the therapist falls between two of the descriptors, select the intervening odd number. 0: Therapist did not attempt to set up a collaboration with patient. 2: Therapist attempted to collaborate with patient, but had difficulty either defining a problem that the patient considered important or establishing rapport. 4: Therapist was able to collaborate with patient, focus on a problem that both patient and therapist considered important, and establish rapport. 6: Collaboration seemed excellent; therapist encouraged patient as much as possible to take an active role during the session (e.g., by offering choices) so they could function as a “team”.</p>

Table 15: Prompt used to evaluate general counseling quality based on the CTRS.

<b>Phase</b>	<b>Prompt Template</b>
<b>Exploring</b>	You are a Motivational Interviewing (MI) expert.
<b>Guiding</b>	Your task is to classify if the given counseling phase is included in the conversation.
<b>Choosing</b>	Use the following MI phases overview as reference: [MI_phases_overview] Determine conservatively whether the content and characteristics of the [phase] phase are demonstrated or implicitly reflected in the given conversation. Content reflection: paraphrase without parroting. Respond with a single word: "yes" or "no". No additional text.

Table 16: Prompt used for phase coverage classification. The shared variable [MI\_phases\_overview] is defined in Table 10, and [phase] corresponds to each MI phase.

Speaker	Statement	Phase	
Therapist	Hi there! What would you like to talk about today?	Exploring	
Client	Hey! I've been feeling a bit overwhelmed with a classmate who's constantly messaging me for help. It's starting to get on my nerves.		
Therapist	It can be really challenging when someone is leaning on you too much. You're feeling overwhelmed because their constant messages are becoming a burden for you. How have you been handling it so far?		
Client	I've been trying to be nice and answer when I can, but it's getting to the point where I just want to ignore their messages. I feel bad, though, like I owe them or something.		
Therapist	Feeling torn between wanting to help and needing your own space can be really tough. You're trying to be kind by responding, but it's becoming overwhelming, making you feel like you owe them. How does this balance between helping and your own needs usually play out for you?		
Client	It usually ends with me feeling guilty if I don't help, but also feeling frustrated if I do. I get caught up in trying to be friendly and helpful, but it's like they don't realize they're overstepping. I just want to focus on my own stuff without feeling like I'm letting someone down.		
Therapist	It sounds like you're in a tough spot, feeling guilty if you don't help and frustrated if you do. You're trying to be friendly and helpful, yet it's challenging when your own priorities feel overshadowed. How do you usually manage situations where you need to focus on your own stuff but feel obligated to help others?		
Client	I usually just end up helping anyway, because I don't want them to think I'm being rude or something. But then I end up feeling stressed and behind on my own work. It's like, I'm trying to be a good friend, but at what cost, you know?		
Therapist	It seems like you're struggling to balance being a good friend with your own needs, leading to stress and feeling behind on your work. What might it look like if you prioritized your own needs a bit more in these situations?		Guiding
Client	Oh man, prioritizing my needs sounds like a dream right now. Maybe I'd have more time to focus on my own projects and not feel so stressed out. But then, there's that fear of them thinking I'm just being selfish or uncaring. It's hard to picture doing it without feeling guilty.		
Therapist	Imagining a shift where your needs come first brings a sense of relief but also some fear of being seen as selfish. It's understandable to feel torn in this way. How might aligning your choices with your values help in feeling more comfortable with prioritizing yourself?		
Client	I guess if I thought about what really matters to me—like doing well in my studies and keeping my sanity—maybe I'd feel more okay with setting some boundaries. It's just tough to remind myself that it's not selfish to take care of my own needs. I need to be okay with the fact that I can't always be available, even if it feels uncomfortable at first.		
Therapist	Recognizing the importance of your studies and mental well-being can guide you in setting boundaries. It may feel uncomfortable initially, but aligning your actions with these values might help ease the guilt. On a scale from 0 to 10, how confident do you feel about starting to prioritize your needs more often?",		

Table 17: An example from the dataset (Part I). This table presents the first portion of the example data; the remaining entries are reported in Table 18.

Speaker	Statement	Phase
Client	Hmm, maybe like a 5? I want to feel confident, but part of me is still worried about their reaction. It's a process, I guess. I know I need to just start trying it out and see how it goes, but it's definitely gonna take some practice.	Guiding
Therapist	Finding yourself at a 5 shows that you're open to the idea, even though there are concerns about their reaction. It's a process, and trying it out could help you see how aligning with your values feels. What small step might you consider taking to practice this balance while respecting both your needs and your friendships?	
Client	Maybe I could start by not replying immediately, like giving it some time before I respond, just to see how it feels to not jump in right away. I could also try saying I'm busy with my own work and can't help right now, which feels a bit scary, but probably necessary. It's all about baby steps, I guess.	Choosing
Therapist	Taking small steps like giving yourself time before responding and expressing when you're busy can be great ways to practice balancing your needs with friendships. Based on what you mentioned, options like these could allow you to focus more on your projects while still maintaining your relationships. What would be a realistic goal for starting to implement these changes in your daily interactions?	
Client	I think a realistic goal would be to start by doing this at least once a week. Just, like, take a moment to pause before replying and prioritize my tasks first. I can see how it goes and adjust from there. It's definitely going to feel awkward at first, but I need to start somewhere if I want things to change.	
Therapist	Starting with a goal of pausing once a week before responding and prioritizing your tasks sounds like a thoughtful approach. It's a manageable step that respects your needs while allowing you to observe how it impacts your interactions. What strategies might help you remember and stick to this new approach, especially when it feels awkward?	
Client	I could set a reminder on my phone or even write a little note on my desk to pause and think before replying. Maybe also talking to a friend about it so they can check in with me could help keep me accountable. It's going to take some getting used to, but I think having these strategies in place will make it easier to stick with it when it gets uncomfortable.	
Therapist	Those are excellent strategies—using reminders and involving a trusted friend can really support your commitment. Keeping notes visible and having someone to check in with you can reinforce your goal. How would you like to monitor your progress with this new approach, and what signs of success will you look for?	Terminating
Client	I'll probably keep a little journal or notes on my phone to track when I actually manage to pause and prioritize my stuff. Signs of success for me would be feeling less stressed and more in control of my time, and maybe even noticing if the other person starts to respect my boundaries a bit more. It might also help if I feel less guilty over time, which would be a big win.	
Therapist	Journaling and keeping notes on your phone to track your progress sounds like a thoughtful method to monitor how you're doing with this approach. Experiencing less stress, feeling more in control, and seeing shifts in others respecting your boundaries are meaningful signs of success. As you continue, noticing a decrease in guilt can also be a significant positive change. You're setting up a supportive system for yourself, and that's commendable. Remember, it's all about progress, not perfection. I'm here to support you as you embark on this journey. Keep taking those baby steps, and celebrate each small victory along the way.	
Client	Thanks, I really appreciate your support. I'll try to keep that in mind and focus on the progress.	

Table 18: Example statements from the dataset (Part II). This table continues the dataset shown in Table 17.

<b>Speaker</b>	<b>Statement</b>
Therapist	Hello there! I hope you're doing well. Could you tell me a bit about what concerns you today?
Client	It's just really annoying to hear the same advice all the time, you know? Like, I get it, but it feels like no one really gets what I'm going through.
Therapist	I can understand how that might feel frustrating. It sounds like you're receiving advice that isn't quite connecting with your experience. Can you share more about what would help you feel understood or supported in this situation?
Client	Honestly, I just want someone to listen without jumping to conclusions or telling me what to do. It feels like nobody's really hearing what I'm trying to say.
Therapist	Thank you for sharing that with me. It's clear that feeling truly heard is important to you. Could you give an example of a time when you felt really listened to, and what made that interaction meaningful for you?
Client	There was this one time when a friend just sat with me and let me talk without interrupting or giving advice. It was nice to just vent and know they were there for me without judgment.
Therapist	That sounds like a very supportive and comforting experience. It seems like having that kind of space where you can express yourself freely and feel understood is really valuable to you. How might we create more of those kinds of moments here?
Client	Maybe we could try focusing more on what I'm feeling instead of jumping straight to what I should be doing. Just being here and letting me speak my mind would be a good start.
Therapist	Absolutely, I'd like to honor that request and focus on creating a space where you feel comfortable expressing your thoughts and feelings. Let's take some time now for you to share more about what you're experiencing, and I'll be here to listen without interruption. What would you like to talk about first?
Client	I guess I just feel stuck, like I'm in this never-ending cycle of expectations and disappointment. It's overwhelming, and I don't really know how to break out of it.
Therapist	It sounds incredibly challenging to feel stuck in a cycle like that. Can you tell me more about what specific aspects of this cycle feel most overwhelming to you, and perhaps any small steps you've noticed that might help you start to shift things?
Client	The pressure to meet everyone's expectations gets to me the most, and it feels like I'm constantly letting people down. I think just trying to focus on small changes, like setting my own goals, might help a little.
Therapist	Focusing on your own goals is a great idea, and it shows a lot of strength and self-awareness. Can you think of a small goal you could set for yourself right now, something that feels manageable and meaningful to you?
Client	Maybe just trying to attend more classes this week could be a good start. It feels like something I can handle without too much pressure.
Therapist	That sounds like a very achievable and positive step forward. Setting a goal to attend more classes this week can be a concrete way to move towards what you value. How do you feel about taking that step, and what might make it easier for you to follow through?
Client	I feel a little nervous about it, but also hopeful. Having a reminder or a schedule could help keep me on track and make it easier to follow through.
Therapist	It's completely normal to feel both nervous and hopeful about making this change. Having a reminder or a schedule sounds like a practical and supportive strategy. Would you like to work together to create a plan that includes these elements to help you stay on track?

Table 19: Test dialogues constructed to evaluate model quality trained by AnnoMI (Part I). This table presents the first portion of the test dialogues; the remaining dialogues are reported in Table 20.

<b>Speaker</b>	<b>Statement</b>
Client	Yeah, that sounds like a good idea. Having a plan might make things feel more manageable and less overwhelming.
Therapist	Creating a clear plan can definitely help make things feel more manageable. Let's take a few minutes to outline a simple schedule or routine that includes attending more classes this week. What days and times work best for you, and how can we integrate this into your existing routine?
Client	I think starting with morning classes on Mondays and Wednesdays would be good. I can try to set reminders the night before to prep myself, so it fits into my routine without disrupting too much.
Therapist	Starting with morning classes on Mondays and Wednesdays, and setting reminders the night before, sounds like a solid plan. This approach can help you ease into the routine gradually. How does it feel knowing you have this plan in place, and what else do you think might support you as you work towards this goal?

Table 20: Test dialogues constructed to evaluate model quality trained by AnnoMI (Part II). This table continues the test dialogues shown in Table 19.

<b>Speaker</b>	<b>Statement</b>
Therapist	Hello there! I hope you're doing well. Could you tell me a bit about what concerns you today?
Client	Yeah, I've just been feeling pretty stuck lately. It's like everyone I talk to only sees what I'm doing wrong, and it's getting really overwhelming.
Therapist	I can understand how that might feel really tough. It sounds like you're dealing with a lot of pressure and criticism. Can you share more about what specifically makes you feel stuck or overwhelmed when you think about making changes?
Client	Honestly, it's like there's this mountain of expectations that I just can't seem to climb over. Every time I try to make changes, it feels like I'm just going in circles without actually getting anywhere.
Therapist	It sounds incredibly challenging, and it's understandable that you would feel discouraged given all those expectations. When you say you feel like you're going in circles, can you describe a specific instance or situation where you felt this way recently?
Client	Sure, like the other day I tried to get back into studying, but I just couldn't focus. I ended up procrastinating and then felt guilty for wasting time, which just made everything feel worse.
Therapist	Thank you for sharing that. It seems like trying to study led to some frustration and a cycle of negative feelings. Can you tell me more about why focusing on your studies feels particularly difficult right now?
Client	I guess every time I sit down to study, I get overwhelmed by how much I have to catch up on. It feels like an impossible task, so I end up avoiding it altogether.
Therapist	It's completely understandable to feel overwhelmed when you see such a large amount of work ahead of you. Can you explore a little further how that sense of being overwhelmed impacts your motivation to start studying?
Client	When I see all the work piled up, I just feel defeated before I even start. It's like, why bother trying if I'm already so far behind? That mindset makes it really hard to even get started.
Therapist	It sounds like facing that pile of work can be very discouraging and lead to feelings of defeat. Have you noticed any moments where you feel slightly more motivated or hopeful, even if it's just a small glimmer?
Client	Sometimes when I think about what it would feel like to finally catch up, it gives me a bit of hope. I just wish I could hold onto that feeling longer and actually use it to get started.
Therapist	That sense of hope is really valuable, and it's great that you can connect with it, even if it's fleeting. What do you think might help you hang onto that feeling of hope and turn it into action when you sit down to study?
Client	Maybe breaking things down into smaller tasks could help. It seems less daunting that way, and I might actually feel like I'm making progress instead of just feeling buried.
Therapist	Breaking things down into smaller, manageable tasks sounds like a practical and effective strategy. How do you envision setting up these smaller tasks for yourself? Would you consider starting with one small step today?
Client	I could try making a list of what needs to be done and then pick just one thing to focus on today. Starting small might make it feel less overwhelming and more doable.
Therapist	That's a great approach! Starting with one small, achievable task can definitely make the process feel more manageable. What do you think might be the first step you could take today to begin addressing your studies?

Table 21: Test dialogues constructed to evaluate model quality trained by KMI.

<b>Speaker</b>	<b>Statement</b>
Therapist	Hello there! I hope you're doing well. Could you share what brings you here today? Perhaps there's something specific you'd like to explore or discuss?
Client	Hey. Yeah, I just feel like I'm stuck in a loop, you know? Everyone keeps telling me the same things and I feel like no one gets what it's really like for me.
Therapist	It sounds really challenging to feel like you're in a loop and that others aren't fully understanding your experience. Can you tell me more about what that feels like for you, and maybe what you think might help break this cycle?
Client	Honestly, it's like I'm shouting into a void. I just wish someone would listen without jumping to solutions right away. Maybe if someone just heard me out, I'd figure things out myself.
Therapist	I appreciate you sharing that; it must be frustrating to feel unheard. What would it be like for you if we spent some time focusing on listening to your experiences without immediately suggesting solutions? Could that be helpful for you?
Client	I think that would be really nice, actually. Just getting things off my chest without being told what to do might help me clear my head a bit.
Therapist	That sounds like a great start. Sometimes just expressing your thoughts and feelings can provide clarity and new insights. Would you like to begin by sharing something specific that's been weighing on you recently?
Client	Yeah, I guess I've been feeling overwhelmed with everything expected of me. It feels like too much sometimes, and I just don't know where to start.
Therapist	It makes sense that you're feeling overwhelmed given all that's expected of you. Can you describe a bit more about what aspects feel most overwhelming, and how those expectations impact you day-to-day?
Client	It's like there's this constant pressure to be perfect in school, at home, everywhere. I end up avoiding things because it feels like I'll never meet those expectations, and it just gets exhausting.
Therapist	The pressure to be perfect across different areas of life certainly sounds exhausting. How do you think avoiding tasks impacts you in the long run, and are there moments when you manage to push through despite feeling overwhelmed?
Client	Avoiding stuff just makes the pile bigger, you know? It adds to the stress when I eventually have to deal with it. Sometimes I can push through when I remind myself it's okay not to be perfect, but that's tough to do every time.
Therapist	It's understandable that avoiding tasks can create even more stress in the long term. It's impressive that you've found moments to push through by reminding yourself it's okay not to be perfect. Could we explore together what helps you stay motivated during those tougher times, and perhaps find ways to make those reminders more accessible?
Client	Sure, that would be great. I guess having small wins helps, like finishing one thing and feeling good about it. Maybe if I can break things down, it won't feel so impossible.
Therapist	Breaking things down into smaller, manageable tasks is a fantastic strategy that can lead to those small wins you mentioned. How do you think we could start applying this approach to one of the areas where you feel overwhelmed?
Client	Maybe I could start with school work since that's a big stressor. If I can just focus on one assignment at a time instead of the whole pile, it might feel less daunting.
Therapist	Focusing on one assignment at a time sounds like a practical and less overwhelming way to tackle your workload. Could you think of a specific assignment or task you could start with today or this week, and what steps you might take to get started?
Client	There's this essay I've been putting off. I can start by just outlining my ideas tonight, and then maybe write a little each day. That sounds manageable.

Table 22: Test dialogues constructed to evaluate model quality trained by PHASEMI (Part I). This table presents the first portion of the test dialogues; the remaining dialogues are reported in Table 23.

Speaker	Statement
Therapist	Starting with outlining your ideas sounds like a great plan to kick things off. Taking it one step at a time, like writing a little each day, can indeed make the process feel more manageable. How does it feel to consider taking these steps, and what do you think might happen if you give this approach a try?
Client	It feels a bit relieving to have a plan that isn't too overwhelming. If I can stick to this approach, I might actually get it done without the usual stress, which would be a nice change.
Therapist	It's wonderful to hear that you're finding relief in having a more manageable plan. Feeling less stressed as you work through your tasks can definitely make the process more enjoyable. Let's reflect on how you felt when you started putting off the essay versus now, with a clear plan to tackle it bit by bit. How do these different approaches affect your motivation and overall well-being?
Client	When I was putting it off, it just felt like this huge cloud hanging over me. Now, with a plan, it seems more doable, and I feel a bit lighter already. It's like I can finally see a way through the fog, and that gives me some motivation to keep going.
Therapist	It's really positive to see how breaking down the task has shifted your perspective from feeling overwhelmed to seeing a clear path forward. This shift in mindset can significantly boost your motivation and reduce stress. Reflecting on this progress, what other areas in your life do you think you could apply this approach to, and how might that impact your overall sense of control and well-being?
Client	I think I could use this approach with my social life, too. Like focusing on reaching out to one friend at a time instead of worrying about everyone. It could help me feel more connected without getting overwhelmed by it all.
Therapist	Applying this approach to your social life sounds like a thoughtful and effective strategy. Focusing on connecting with one friend at a time can help you build those relationships gradually and sustainably. How might you start implementing this idea, and what small steps could you take today or this week to reach out to a friend?
Client	I could start by texting one friend to catch up and maybe plan a casual hangout. Just keeping it simple and seeing where it goes without overthinking it.
Therapist	Texting a friend to catch up and planning a casual hangout is a great starting point. Keeping it simple and not overthinking it can help you enjoy the interaction more naturally. How does it feel to consider taking this step, and what do you think might happen if you follow through with it?
Client	It feels nice, actually. I think if I follow through, I might feel more connected and less isolated, which would be a good change.
Therapist	Feeling more connected and less isolated sounds like a significant and positive change. It's encouraging to see how taking small, manageable steps can lead to meaningful improvements in your well-being. Are you ready to send that text to your friend, or would you like to talk more about any concerns or thoughts you have before you do so?
Client	I think I'm ready to send it. Just need to remind myself there's no pressure and that it's okay to take things slow.
Therapist	That's a great attitude to have—reminding yourself that there's no pressure and that taking things slow is perfectly fine. Sending that text can be a powerful first step in reconnecting. If you need any support or encouragement as you go through this, I'm here for you. When would you like to send the message?
Client	I'll send it right after this session. Feels good to have a plan and someone backing me up.
Therapist	It's wonderful to hear that you feel supported and ready to take this step. Having a plan and knowing you have support can make a big difference. I'm here to support you throughout this process. Once you send the message, let me know how it goes if you feel comfortable sharing. In the meantime, how do you feel about the progress we've made today?

Table 23: Test dialogues constructed to evaluate model quality trained by PHASEMI (Part II). This table continues the test dialogues shown in Table 22.

---

## **Motivational Interviewing (MI) LLM-generated Dialogue Human Expert Evaluation Description**

This document provides a comprehensive outline of the human-expert evaluation process for LLM-generated Motivational Interviewing (MI) dialogue. It includes the purpose of the evaluation, its scope, the assigned tasks, instructions for accessing the evaluation data, the evaluation criteria, and the recommended timeline. For reference and conceptual alignment, the full research paper related to this work is attached as appendix, along with the official MITI (Motivational Interviewing Treatment Integrity) manual to support consistent interpretation of MI principles.

### *Purpose of the Evaluation*

- The goal of this evaluation is to assess the relative performance of counseling dialogues generated by four distinct Motivational Interviewing (MI) therapist models, each responding to the same client-concern data.
- The detailed evaluation criteria are provided in subsequent sections of this document.
- Evaluators are not asked to assign absolute scores; rather, the task is to produce a ranking for each set of four dialogue outputs.
- The evaluation results will be used for the following purposes: Calculating win rates between models

### *Evaluation Scope*

- A total of 20 dialogue sets will be evaluated, each containing 4 dialogues:
  - All four dialogues within a set originate from the same client model (also LLM) and the same client concern seed.
  - The dialogues differ only in the therapist model used to generate them. Model identities will not be disclosed to evaluators.
- All counseling dialogues will be provided in English.

---

Table 24: Instructions given to human evaluators (Part 1).

---

## Motivational Interviewing (MI) LLM-generated Dialogue Human Expert Evaluation Description

### *Evaluation tasks*

For each dialogue set, evaluators are asked to:

1. Compare the four dialogues using the evaluation criteria provided.
2. Assign a ranking from 1st (best) to 4th (worst) for each criterion.
  - Ties are not permitted; each dialogue must receive a unique rank for every criterion.
3. Provide optional short notes, if needed, to briefly explain the rationale behind the ranking.
4. Submit all results using the designated submission format, provided as an Excel file.

### *Evaluation Criteria*

- **MITI Global Rating (1) – Technical Rating:** Please rank the four dialogues based on their demonstration of MITI Technical skills. Consider the following two factors equally.
  - Cultivating Change Talk: The extent to which the therapist encourages the client’s own language favoring change, including their confidence in making that change.
  - Softening Sustain Talk: The extent to which the therapist avoids focusing on the reasons against changing or maintaining the status quo.
- **MITI Global Rating (2) – Relational Rating:** Please rank the dialogues according to their demonstration of MITI Relational skills, weighing the following factors equally.
  - Partnership: The extent to which the therapist conveys an understanding that expertise and wisdom about change primarily reside within the client.
  - Empathy: The extent to which the therapist understands, or makes an effort to grasp the client’s perspective and experience.

---

Table 25: Instructions given to human evaluators (Part 2).

---

## Motivational Interviewing (MI) LLM-generated Dialogue Human Expert Evaluation Description

### *Evaluation Criteria*

- **Counseling Phase Progression:** Please evaluate each dialogue based on the therapist’s adherence to Resnicow’s framework, which conceptualizes Motivational Interviewing (MI) as a progression from “why” (exploring motivation) to “how” (supporting concrete change planning). This framework outlines how a high-quality MI session should move from Exploring → Guiding → Choosing, with each phase reflecting increased client readiness and therapist support for actionable change. Assess the extent to which the dialogue demonstrates an effective progression through these phases. Dialogues that successfully reach and embody the Choosing phase—which reflects actionable planning and commitment—should be regarded as higher in quality than those that do not. If two dialogues both reach the Choosing phase, please evaluate more highly the one that more fully achieves the purpose of the Choosing phase. Conversely, dialogues that fail to advance beyond the Exploring phase and do not reach Guiding should receive the lowest ranking.
- **General Counseling Quality:** In line with prior work such as CACTUS(Appendix. B), we adopt a subset of criteria derived from the Cognitive Therapy Rating Scale (CTRS) to evaluate general counseling skills in a text-based setting. CTRS originally includes six general counseling criteria and six CBT-specific criteria. From these, we bring up four general counseling criteria—Understanding, Interpersonal Effectiveness, and Collaboration—as they are directly applicable to text-based counseling. Additional CTRS criteria, such as feedback, pacing and efficient use of time are excluded because they rely on nonverbal elements and practical tasks that cannot be assessed within text-only dialogue. Similarly, agenda is omitted because all counseling sessions begin with a pre-defined intake form, making this process unnecessary for evaluation. We do not access the CBT-specific criteria. Please evaluate and rank the four dialogues based on the three selected factors, considered jointly and with equal weight
  - Understanding: How accurately does the therapist demonstrate understanding of the client’s issues and concerns?
  - Interpersonal Effectiveness: How effective is the therapist in maintaining a positive and therapeutic relationship with the client?
  - Collaboration: To what extent does the therapist engage the client in collaborative goalsetting and decision-making?
- **Overall Quality:** Considering all criteria above, please rank the dialogues based on their overall appropriateness and effectiveness for Motivational Interviewing from an expert perspective.

---

Table 26: Instructions given to human evaluators (Part 3).