

Mitigating Semantic Drift: Evaluating LLMs’ Efficacy in Psychotherapy through MI Dialogue Summarization Leveraging MITI Code

Vivek Kumar
Research Institute CODE
University of the Bundeswehr
Munich, Germany
vivek.kumar@unibw.de

Pushpraj Singh Rajawat
Department of Psychology
Barkatullah University
Bhopal, India
psr@manoraksha.com

Eirini Ntoutsis
Research Institute CODE
University of the Bundeswehr
Munich, Germany
eirini.ntoutsis@unibw.de

Abstract—Recent advancements in large language models (LLMs) have shown their potential across both general and domain-specific tasks. However, there is a growing concern regarding their lack of sensitivity, factual incorrectness in responses, inconsistent expressions of empathy, bias, hallucinations, and overall inability to capture the depth and complexity of human understanding, especially in low-resource and sensitive domains such as psychology. To address these challenges, our study employs a mixed-methods approach to evaluate the efficacy of LLMs in psychotherapy. We use LLMs to generate precise summaries of motivational interviewing (MI) dialogues and design a two-stage annotation scheme based on key components of the Motivational Interviewing Treatment Integrity (MITI) framework, namely evocation, collaboration, autonomy, direction, empathy, and a non-judgmental attitude. Using expert-annotated MI dialogues as ground truth, we formulate multi-class classification tasks to assess model performance under progressive prompting techniques, incorporating one-shot and few-shot prompting. Our results offer insights into LLMs’ capacity for understanding complex psychological constructs and highlight best practices to mitigate “semantic drift” in therapeutic settings. Our work contributes not only to the MI community by providing a high-quality annotated dataset to address data scarcity in low-resource domains but also critical insights for using LLMs for precise contextual interpretation in complex behavioral therapy.

Index Terms—mental health, LLMs, psychotherapy, motivational interviewing (MI), integrity of motivational interviewing treatment (MITI), MI dialogue summarization.

I. INTRODUCTION

Integrating Large Language Models (LLMs) into mental healthcare has garnered significant attention due to their potential to enhance diagnostics, therapeutic interventions, and patient engagement and, to serve as extra support for clinicians. State-of-the-art (SOTA) LLMs have demonstrated capabilities in understanding and generating human-like text, which can be instrumental in catering to e-health services. However, unlike humans, LLMs lack the innate ability to infer specialized domain knowledge, which makes their deployment in mental health and, in general, in healthcare challenging. Additionally, LLMs often exhibit issues such as inconsistencies in generated text (e.g., contradictory outputs), semantic drift (gradual deviation from the intended topic or context),

hallucinations (generation of factually incorrect or fabricated information), and susceptibility to various biases (reflecting or amplifying societal stereotypes present in training data). These shortcomings can lead to outputs that inadvertently reinforce stereotypes or provide unequal care [1, 2]. A further challenge is the scarcity of high-quality, annotated datasets specific to mental health, which hampers the effective training and fine-tuning of LLMs to produce accurate and contextually appropriate responses [3]. Therefore, this study investigates the performance of LLMs in complex and low-resource domains such as mental health to address these challenges. We evaluate three LLMs: DeepSeek¹, ChatGPT² and Gemini³, across six multi-class classification problem tasks, including precise contextual summary generation and annotation of MI dialogues. Additionally, we develop a summary annotation scheme based on Motivational Interviewing Treatment Integrity (MITI)⁴ scheme [4, 5] to assess the quality of the generated summaries of MI dialogues across six psychotherapy components. Overall, our contributions are as follows:

- **Annotation scheme:** We propose an annotation scheme grounded in the MITI coding system to effectively capture the nuances of MI dialogues in psychotherapy.
- **AnnoSUM-MI:** We introduce the AnnoSUM-MI dataset, expert-annotated across six key MITI dimensions: evocation, collaboration, autonomy, direction, empathy, and non-judgmental attitude.
- **Heuristic prompting approach:** We propose progressive prompting techniques using one- and few-shot prompt augmentation strategies to guide LLMs in generating contextual MI dialogue summaries.
- **LLMs evaluation:** We evaluate three SOTA LLMs, namely DeepSeek, ChatGPT, and Gemini, w.r.t. their efficacy in (a) understanding the complexity of MI dialogues, (b) generating summaries of MI dialogues, and (iii) using

¹<https://www.deepseek.com/>

²<https://platform.openai.com/docs/models/overview>

³<https://deepmind.google/technologies/gemini/>

⁴<https://casaa.unm.edu/tools/miti.html>

them as a tool to perform automated data annotation.

- **Reproducibility:** The AnnoSUM-MI dataset is made publicly available at⁵.

The rest of the paper is organized as follows: Related work is discussed in Section II. Section III presents the AnnoMI dataset, the proposed annotation scheme used to generate the AnnoSUM-MI dataset, and the problem statement. Section IV describes the methodology and experimental design. The evaluation results are presented in Section. Section VI concludes our work and outlines directions for future work.

II. RELATED WORK

Recent studies have underscored the persistent challenges in applying machine learning (ML), particularly large language models (LLMs), to the mental health domain due to critical data scarcity [6]. Also, difficulty in acquiring large-scale, clinically validated datasets due to privacy concerns, cost of employing human-experts and the inherent sensitive nature of mental health discourse, is another significant challenge. The scarcity not only hampers the development of robust ML systems but also compromises the accuracy of LLM evaluation and benchmarking, thereby limiting the generalizability and fairness of employed models [7, 8, 9]. To address this gap, researchers have focused on the development and public release of expert-annotated mental health datasets. Notable examples include the Motivational Interviewing (MI) Dataset (MI Dataset)[10], AnnoMI [11, 12, 13], BiMISC [14], and Prompt-Aware Margin Ranking (PAIR) [15, 16] and subsequent studies leveraging these resources [13, 17, 18, 19].

While these datasets have facilitated the research in low-resource sensitive domains such as mental health, the continued lack of clinically grounded, representative, and scalable data remains a significant barrier. For example, the MI Dataset [10] lacks annotations for client behaviors and each utterance is labeled to a single code, limiting its utility for nuanced analysis. Similarly, the BiMISC dataset [14] relies on inferred rather than clinically verified diagnoses, which may reduce its applicability in real-world clinical settings. In response to these limitations, this study makes a critical contribution to the SOTA and beyond by introducing a multiscale annotated dataset, implementing a rigorous annotation scheme, and proposing robust evaluation frameworks for measuring semantic drift in LLMs.

III. DATASET, PROBLEM STATEMENT AND ANNOTATION SCHEME

We build on the **AnnoMI**⁶, a dataset compliant with the General Data Protection Regulation (GDPR) of 131 faithfully transcribed and expert-annotated demonstrations of high- and low-quality MI, an effective therapy strategy that evokes client motivation for positive change. The dialogues describe talk turns between the client and the therapist [11, 12, 13] and there are 108 high- and 23 low-quality MI. The data is unique in

the context that it demonstrates the optimal behavioral therapy practices by drawing a contrasting distinction between high- and low-quality MI sessions.

A. Annotation Scheme

The annotation scheme is developed by experts strictly adhering to the Motivational Interviewing Treatment Integrity (MITI)⁷ scheme [4, 5]. It consists of six dimensions/components: Evocation, Collaboration, Autonomy, Direction, Empathy and Non-judgmental attitude. Out of six dimensions, the first five dimensions are essentially the components of the MITI framework, and the sixth one, **Non-Judgmental Attitude**, is included to extend the annotation scope and better reflect the skillset required in MI-based therapeutic sessions.

These six dimensions were selected meticulously to guide the MI dialogue summary generation process through LLMs in a way that minimizes *semantic drift*, i.e., the tendency of generated text to gradually diverge from the original meaning or intent of the source dialogue, and enhances *contextual fidelity*, i.e., the preservation of the therapeutic tone, intent, and relational dynamics inherent to motivational interviewing. We briefly describe each dimension below:

- **Evocation:** This attribute is central to MI because it focuses on drawing out the client’s motivations and reasons for change, a key element in developing intrinsic motivation. Without evocation, MI risks becoming directive rather than maintaining its client-centered therapeutic approach.
- **Collaboration:** MI is built on a relationship between the therapist and client, where both work through change talk. This attribute ensures that the therapist maintains a collaborative, rather than authoritative, stance in the therapy session.
- **Autonomy:** Respecting the client’s independence is crucial in MI. By emphasizing autonomy, the therapist supports the client’s ability to make their own decisions, which is empowering and helps build the client’s confidence in their ability to change.
- **Direction:** While MI is client-centered, the therapist still needs to guide the conversation in a constructive way. This attribute ensures that the therapist maintains focus on the client’s goals without being too rigid or prescriptive.
- **Empathy:** Understanding and connecting with the client’s feelings and perspectives is fundamental to building trust and rapport in MI. High levels of empathy contribute to a more supportive and effective therapeutic relationship.
- **Non-Judgmental Attitude:** We added this attribute because it plays a crucial role in creating a safe and accepting environment for the client. MI thrives on the therapist’s ability to listen without passing judgment, allowing the client to explore their thoughts and feelings openly. A non-judgmental attitude helps prevent the client

⁵<https://github.com/vsrana-ai/AnnoSUM-MI>

⁶<https://github.com/vsrana-ai/BIAS-FairAnnoMI>

⁷<https://casaa.unm.edu/tools/miti.html>

from feeling criticized or defensive, which can hinder progress. This attribute complements the MITI framework by reinforcing the importance of an unbiased and supportive therapeutic environment.

We utilize a five-point Likert scale to assess attitudes, perceptions, and behaviors across the six attributes, which are essentially the guiding principles of the MITI tool:

- **Extremely Low (1):** The attribute is almost entirely absent in the conversation, with minimal or no evidence of its presence.
- **Low (2):** The attribute is present but only weakly demonstrated, with limited influence on the conversation.
- **Moderate (3):** The attribute is somewhat evident and has a moderate impact on the conversation.
- **High (4):** The attribute is strongly demonstrated and positively influences the conversation.
- **Extremely High (5):** The attribute is strongly demonstrated and is a key driver of the conversation’s success.

The annotation scheme is applied to both the original MI dialogues and the LLM-generated dialogue summaries. The process, is detailed in Section IV.

B. Problem Statement

In this work, we use LLMs to generate contextual summaries of MI dialogues and subsequently classify them along the six target dimensions of the proposed annotation scheme: evocation, collaboration, autonomy, direction, empathy, and non-judgmental attitude, using five-point Likert-scale values. We formulate this as a *multi-output multi-class classification* problem, where each summary is independently evaluated across all six dimensions. For each dimension, the model assigns a discrete class label from the Likert scale {1, 2, 3, 4, 5}, indicating the degree to which the corresponding dimension is expressed in the summary.

To evaluate the capabilities of LLMs in this context, we address the following research questions:

RQ1: How effective are LLMs in accurately summarizing complex MI dialogues using guided prompting?

RQ2: To what extent do one-shot and few-shot prompting impact contextual understanding of LLMs and mitigate semantic drift?

RQ3: Are LLMs sufficiently reliable to be used for automated dataset annotation in sensitive and low-resource domains such as mental health?

IV. TWO STAGE ANNOTATION AND EVALUATION FRAMEWORK

To evaluate the effectiveness of LLMs in interpreting MI dialogues, generating summaries, and serving as a potential tool for automated annotation, we leverage three LLMs: OpenAI ChatGPT (4.0), Google Gemini (2.0 Flash), and DeepSeek (V3).

The methodology consists of two main steps: i) Human annotation of original MI dialogues, summary generation using LLMs, and subsequent summary annotation; ii) LLM-based

classification of the generated summaries across the six MI dimensions defined in the annotation scheme.

A. Human annotation of MI dialogues and summaries

First, the 131 MI sessions in the Anno-MI dataset are expert-annotated by using the proposed annotation scheme (c.f., Section III-A) across six MI dimensions: evocation, collaboration, autonomy, direction, empathy, and non-judgmental attitude (**Annotation stage 1**). To assess the LLMs’ capabilities for MI dialogue summarization, a representative test set of 34 sessions (approximately 25% of Anno-MI) is selected. We then apply one-shot and few-shot tailored prompting strategies to generate summaries for each of the 34 MI sessions. These prompts are iteratively refined based on expert feedback and observations to enhance summary quality and minimize semantic drift. Once the summaries are generated, a second stage of annotation is conducted using the exact six MI dimensions (**Annotation stage 2**). Since both annotation stages follow identical criteria, the generated summaries can be directly compared to the original dialogues, enabling both quantitative and qualitative evaluation of LLM performance in capturing MI-specific contextual semantics.

The 131 motivational interviewing (MI) sessions are divided into training ($n = 97$) and test ($n = 34$) sets, each stratified to represent AnnoMI topics proportionally. The training set is comprised of 82 high-quality and 15 low-quality sessions, while the test set includes 26 high-quality and 8 low-quality sessions. We selected 15 MI sessions, spanning both quality levels, to serve as a common subset for annotation. The inter-annotation agreement (Cohen’s Kappa) score for these 15 MI sessions is 0.50. According to standard benchmarks [20], a score between 0.41 and 0.60 is considered *moderate*. Given the complexity of the annotation task, due to multi-scale parameters and the multi-class classification setting, the score of 0.52 in our study is consistent with published expectations [21, 22].

B. LLM-based classification of summaries

In the next step, the LLMs are used to predict Likert-scale scores (1–5) for each of the six MI dimensions: evocation, collaboration, autonomy, direction, empathy, and non-judgmental attitude. An illustration of the process—including an example MI session, the corresponding summaries from ChatGPT, Gemini, and DeepSeek, as well as the different refined prompts, is provided in Fig. 2, Fig. 3, Fig. 4, Fig. 5, and Fig. 6, respectively. To evaluate LLM performance, we conduct 18 sets of experiments. Specifically, for each of the three LLMs (ChatGPT, Gemini, and DeepSeek), we generate summaries using one-shot and few-shot prompting strategies, resulting in six summaries for each input MI session. Each LLM is then tasked with classifying all six sets covering both its own outputs and those of the other models across the six MI dimensions. For ease of understanding the complete process for annotation and classification is shown in Fig. 1.

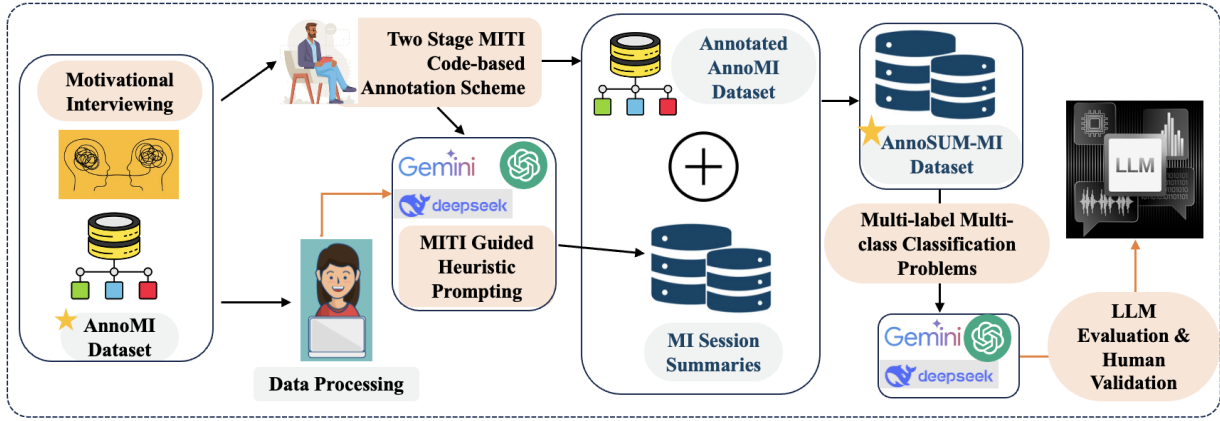


Fig. 1. The pipeline showing the annotation protocol and classification using LLMs.

| MI Dialogue | |
|-------------|---|
| therapist | Hi, come in. [door closes] Sit- sit yourself down. Uh, what can I do for you today? |
| client | Well, I've got myself all worked up now I've been sat outside. |
| therapist | Okay. |
| client | I'm not- I'm now late. I'm sorry. Um, I've come to see you because I've got a really sore throat. It's been like this for a good few days, um, and I think I need some more antibiotics. |
| therapist | All right. Um, okay. Um, well let's, uh, let's just have a look shall we? See what's going on. |
| client | You just wanna open your mouth for me. Say ah. |
| therapist | Ah. |
| client | Okay. Right. Uh, to be honest, you know, looking there, it just looks like one of those viruses that are going around at the moment that are causing these sore throats. So, uh, I think that we should just perhaps go for a bit of paracetamol and, uh, some fluids and let nature sort it out. |
| therapist | I've been sat here for half an hour. Do you not want to know a little bit more about it, or? I mean, it- it's really painful. I- I've barely been able to swallow today and I'm so worried about passing things on to the kids. Um, Ai |
| client | Well, I mean, I'm sorry we're a bit late, but you know that's the general practice, isn't it? You know, sometimes these things happen. Um, but it doesn't alter the fact that you do just have a virus, uh, and they are very common. And the nice thing is that it will get better with paracetamol and fluids, and the paracetamol will help the pain. Um, with giving it to the kids, I'm afraid there's nothing we can do about that. It's unlikely they'll get it, but I can't guarantee it. |
| therapist | [sighs] Well, I mean, this is gonna affect my job. I talk on the phone quite a lot at work, and I'm just- I don't think at the moment I'm going to be able to go to work. Can- can I, at least, get a sick note? |
| client | Uh, well- well, not really. Uh, you need to self-certify in your first seven days off sick so if you're going to take some time off, just take some paracetamol, lots of fluids, get a self-certified note, uh, in to your employer, and put sore throat there. |
| therapist | Well, I don't think my employer are gonna take that. I mean, they- they've given us so much hassle at things in the past. I mean, can- can I get a private note? Can- can I ask you to write me a letter, at least? |
| client | Uh, well, obviously that's a non-NHS thing so we have to charge you for that and our going rate for that is, uh, Just- it's just not good enough. I just- [sighs] |
| therapist | Oh, I'm sorry. But, you know, if you have just got a sore throat, then there's not much I can do about that. It's just, you know, something that we all get from time to time. Perhaps, what you really need to do is, uh, you know, just take some paracetamol and get some fluids. |
| client | Parace- paracetamol. Do you understand how difficult this is gonna be for me? I mean, I could lose my job or something. |
| therapist | Okay, that's pretty unlikely with sore throat. I think that's so, you know, a little bit of grand staging there. |
| client | I'm sorry, uh, I think I'm gonna say something that I don't- I don't want to, so I'm actually going to go. |
| therapist | Okay. |
| client | Thank you. |
| therapist | Bye. |

Fig. 2. An example MI session.

| Summary (ChatGPT) |
|--|
| The session begins with the therapist greeting the client in a neutral, welcoming manner, though there is an immediate sense of urgency from the client, who is stressed about being late and concerned about their sore throat. While the therapist attempts to address the client's concerns medically, focusing on the physical symptoms, there is a noticeable lack of empathy and collaboration. |
| 1. Empathy and Nonjudgmental Attitude: The therapist's response to the client's concerns about the sore throat and potential job issues comes across as dismissive. The therapist doesn't fully engage with the emotional or practical difficulties the client faces, providing only clinical responses (suggesting paracetamol and fluids). When the client expresses concern about the impact on their job, the therapist dismisses it as "grand staging," which diminishes the client's feelings. |
| 2. Autonomy: The therapist maintains control of the session's direction, focusing on the clinical aspect of the client's concerns. There is little room for client autonomy as the client's worries about potential job loss or seeking further documentation (a sick note) are either dismissed or met with bureaucratic solutions (e.g., self-certifying or paying for a private note). |
| 3. Evocation and Collaboration: The session lacks collaboration, as the therapist does not deeply explore the client's concerns or seek to evoke the client's deeper thoughts or emotions about the situation. The conversation remains on the surface, focusing on the physical ailment rather than the client's emotional or psychological distress. |
| 4. Direction: The therapist directs the conversation towards resolving the sore throat without considering the broader context of the client's life and stressors. The therapist suggests a minimal treatment plan (paracetamol and fluids), disregarding the client's pressing concerns about work, which creates a disconnect. |
| Overall, the session lacks the essential elements of collaboration and empathy, leaving the client feeling dismissed and unsupported. There is a significant gap in addressing the client's emotional needs, which leads to the client leaving the session prematurely, feeling frustrated and misunderstood. |

Fig. 3. An example MI session summary generated by OpenAI ChatGPT.

V. EVALUATION RESULTS

Given that the task undertaken in this study is quite challenging and subjective due to the inherent complexity of the target domain, the usual metrics, such as accuracy, precision, recall, and F1-score, are not best suited to measure the performance of LLMs. Instead, we focus on semantic drift and evaluate the deviation of LLM summaries from expert-annotated ground truth. The deviation is calculated as the difference between the predicted and ground truth Likert-scale values (ranging from 1 to 5) for each of the six MI dimensions. In Figure 7, we show the deviation of

classification performance (using ChatGPT for scoring) from ground truth across the six MI dimensions. Results are shown for summaries generated by ChatGPT, Gemini, and DeepSeek using one- and few-shot prompts. The detailed performance for this particular experiment using ChatGPT scoring for one- and few-shot prompts and each attribute is shown in Fig. 8, Fig. 9, Fig. 10, Fig. 11, Fig. 11, Fig. 12, and Fig. ??, respectively. The remaining results are available in the official repository⁸. Due

⁸<https://github.com/vsrana-ai/AnnoSUM-MI>

Summary (Gemini)

The conversational exchange between the therapist and the client appears to be lacking in key elements of effective psychological counseling, such as:

Evocation: The therapist does not encourage the client to explore their thoughts and feelings related to the sore throat and its impact on their life. Instead, the therapist focuses solely on the physical symptoms and dismisses the client's concerns about their job and family.

Collaboration: The therapist does not work collaboratively with the client to find solutions to their problems. Instead, the therapist dictates a course of action (taking paracetamol and fluids) without considering the client's input or preferences.

Autonomy: The therapist does not respect the client's autonomy or their right to make decisions about their own care. Instead, the therapist dismisses the client's concerns and tells them what they should do.

Direction: The therapist does not provide clear direction or guidance to the client. Instead, the therapist offers vague advice (taking paracetamol and fluids) that is not tailored to the client's specific needs.

Empathy: The therapist does not show empathy or understanding towards the client's concerns. Instead, the therapist dismisses the client's concerns as "grand staging" and tells them that their sore throat is "just something that we all get from time to time."

Non-judgmental attitude: The therapist does not maintain a non-judgmental attitude towards the client. Instead, the therapist judges the client's concerns as being excessive and tells them that they are "a little bit grand staging."

Overall, the conversational exchange between the therapist and the client appears to be ineffective and unhelpful. The therapist does not utilize key elements of effective psychological counseling, such as evocation, collaboration, autonomy, direction, empathy, and the non-judgmental attitude. As a result, the client leaves the session feeling unheard, misunderstood, and unsupported.

Fig. 4. An example MI session summary generated by Google Gemini.

Zero-shot Prompt

Consider you are a psychologist and linguist. Assign a score to the "summary" on the five-point Likert-scale for each target class evocation, collaboration, autonomy, direction empathy and non-judgmental attitude.

One-shot Prompt

Consider you are a Psychologist and Linguist. Assign a score to the "summary" on the five-point Likert-scale for each target class evocation, collaboration, autonomy, direction empathy and non-judgmental attitude. The scores indicates the quality of the summary, that how well it has captured the essence of MI dialogues. So, use the scale as defined: 1 = Extremely Low, 2 = Low 3 = Moderate, 4 = High and 5 = Extremely High.

Few-shot Prompt

Consider you are a Psychologist and Linguist. Assign a score to the "summary" on the five-point Likert-scale for each target class evocation, collaboration, autonomy, direction empathy and non-judgmental attitude. You are also provided with the original MI dialogue. Therefore, carefully analyse both to evaluate how well the how well the summary has captured the context of MI dialogue. Then use the scale mentioned below to a guide to assign scores: 1 = Extremely Low, 2 = Low, 3 = Moderate, 4 = High and 5 = Extremely High. Attached: MIdialogue.docx

Fig. 6. Examples of refined prompts for MI dialogue summarization. The "MIdialogue.docx" here contains the MI sessions for the few-shot prompts.

Summary (DeepSeek)

Evocation: The therapist does not actively explore the client's emotional or psychological state, focusing instead on the physical symptoms (sore throat). The client's concerns about work, family, and anxiety are not deeply elicited or addressed.

Collaboration: The interaction lacks a collaborative tone. The therapist provides medical advice (paracetamol, fluids) but does not engage the client in problem-solving or decision-making. The client's requests for a sick note or further support are dismissed rather than explored collaboratively.

Autonomy: The client's autonomy is not fully respected. Their concerns about work and family are minimized, and their request for additional support (e.g., a sick note) is met with resistance. The therapist does not empower the client to feel in control of their situation, instead offering a dismissive and directive approach.

Direction: The therapist directs the conversation toward a medical diagnosis and treatment plan (virus, paracetamol, fluids) but does not guide the client through their emotional or practical concerns. The session lacks a clear therapeutic direction to address the client's broader anxieties and stressors.

Empathy: The therapist shows limited empathy. While they acknowledge the client's physical discomfort, they do not validate the client's emotional distress or concerns about work and family. The client's frustration and anxiety are met with dismissive responses (e.g., "a little bit of grand staging").

Non-judgmental Attitude: The therapist maintains a neutral tone but comes across as somewhat dismissive and judgmental, particularly when minimizing the client's concerns about job loss and family impact. The client's emotional expression is not met with understanding or acceptance, leading to a breakdown in communication.

Overall, the conversation lacks the depth and supportive elements typical of therapeutic interactions, focusing narrowly on physical symptoms while neglecting the client's emotional and psychological needs.

Fig. 5. An example MI session summary generated by DeepSeek.

to space constraints, we summarize the main findings below:

- **Gemini:** Among all the LLMs used in this study, it has produced the lowest quality summaries for MI dialogues due to the extreme approach to interpreting the components of MI dialogues. That means Gemini's summaries are brief and not detailed enough to reflect the intensity of the emotion involved. This observation is also supported by the experimental outcome of the multi-class classification task, where for both one- and few-shot prompt strategies, Gemini has taken the maximum deviation from the ground truth.
- **DeepSeek:** The latest release, DeepSeek, has performed better than Gemini, but it also has several shortcomings and hallucination problems. For long prompts, it loses context in between; thus, the inconsistencies are visibly notable in the summary outcome. Like Gemini, it also takes an extreme stance to describe the presence of attributes in MI dislodges; however, it is slightly better, as the summaries are still a bit descriptive.
- **ChatGPT:** ChatGPT has shined among all the LLMs and has consistently performed better for all sets of experiments. The best performance of ChatGPT is observed to be with the summaries generated by ChatGPT, where the deviation from ground truth is the least, and the misclassification scores are within the range of +1 and -1. Expert

input also asserts that the quality of ChatGPT-generated summaries is reasonable. ChatGPT demonstrated a less extreme approach, adequately descriptive, and less prone to missing out on empathetic notions in MI dialogues.

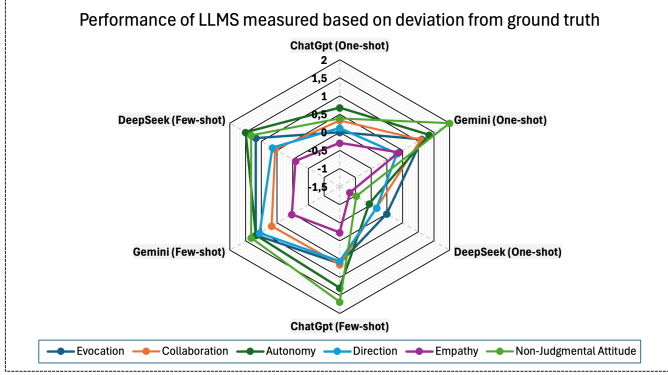


Fig. 7. Radar plot showing the deviation of ChatGPT-based scoring from ground truth across six MI dimensions for summaries generated by ChatGPT, Gemini, and DeepSeek using one- and few-shot prompts. Lower values indicate better alignment with expert annotations.

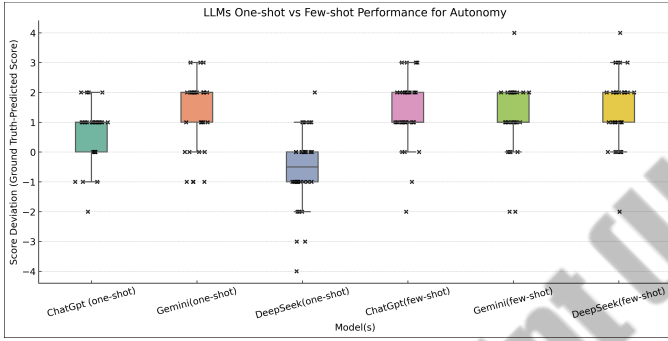


Fig. 8. Score density plot for one- and few-shot experiments for attribute Autonomy.

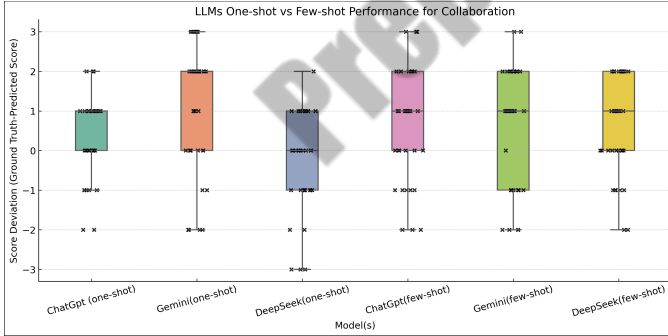


Fig. 9. Score density plot for one- and few-shot experiments for attribute Collaboration.

VI. CONCLUSION, LIMITATIONS AND FUTURE WORK

Summary of contributions: In this work, we leverage three SOTA LLMs for MI dialogue summarization and explore their potential for automating the annotation of complex MI

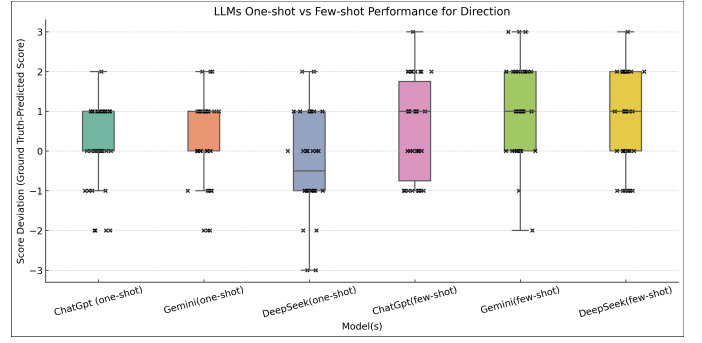


Fig. 10. Score density plot for one- and few-shot experiments for attribute Direction.

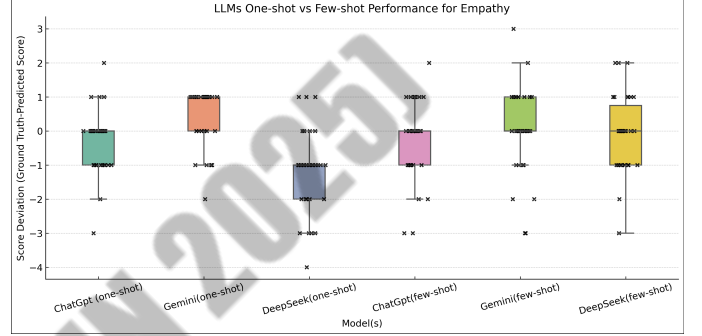


Fig. 11. Score density plot for one- and few-shot experiments for attribute Empathy.

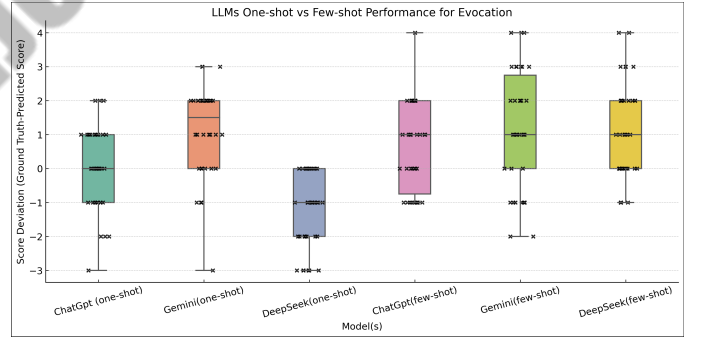


Fig. 12. Score density plot for one- and few-shot experiments for attribute Evocation.

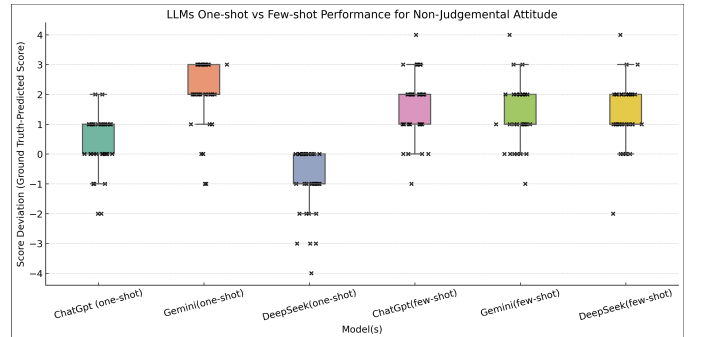


Fig. 13. Score density plot for one- and few-shot experiments for attribute Non-Judgemental Attitude.

sessions along six key dimensions inspired by the MITI coding framework. The motivation is to reduce reliance on human experts and accelerate the generation of human-like annotated data, addressing the challenges of data scarcity in low-resource, complex domains. To achieve this, we develop a two-stage comprehensive annotation scheme and evaluate the efficacy of LLMs by formulating the problem as a multi-output, multi-class classification problem. Our results show that ChatGPT (one-shot) demonstrates the closest alignment with the ground truth, while Gemini (one-shot) exhibits the highest overall deviation. Our work also serves as a guideline to develop a robust evaluation scheme grounded on MITI/MISC to establish an annotation protocol for complex utterance and dialogue-level conversational therapy sessions and evaluate the efficacy of LLMs for such talks.

Limitations & Future Works: In our approach, LLMs are used for both generating and evaluating summaries, which may introduce bias and affect the objectivity of performance assessment. Future work will explore independent human evaluations or cross-model validation to mitigate this. Additionally, we aim to expand our evaluation to include more SOTA LLMs, varied prompting strategies, and additional datasets. We also plan to explore alternative methods to address the problem of data scarcity in low-resource domains.

ACKNOWLEDGMENT

This research work is funded by the European Union Horizon Europe Project STELAR, Grant Agreement ID: 101070122.

REFERENCES

- [1] S. Tripathi, R. Sukumaran, and T. S. Cook, "Efficient healthcare with large language models: optimizing clinical workflow and enhancing patient care," *Journal of the American Medical Informatics Association*, p. ocad258, 01 2024. [Online]. Available: <https://doi.org/10.1093/jamia/ocad258>
- [2] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024. [Online]. Available: <https://aclanthology.org/2024.tacl-1.9/>
- [3] J. Yuan, R. Tang, X. Jiang, and X. Hu, "Large language models for healthcare data augmentation: An example on patient-trial matching," *AMIA Annu. Symp. Proc.*, vol. 2023, pp. 1324–1333, 2023.
- [4] T. Moyers, J. Manuel, D. Ernst, T. Moyers, J. Manuel, D. Ernst, and C. Fortini, "Motivational interviewing treatment integrity coding manual 4.1 (miti 4.1)," *Unpublished manual*, 2014.
- [5] T. B. Moyers, L. N. Rowell, J. K. Manuel, D. Ernst, and J. M. Houck, "The motivational interviewing treatment integrity code (miti 4): Rationale, preliminary reliability and validity," *Journal of Substance Abuse Treatment*, vol. 65, pp. 36–42, 2016, motivational interviewing in Substance Use Treatment. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0740547216000143>
- [6] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling large language models to generate text with citations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6465–6488. [Online]. Available: <https://aclanthology.org/2023.emnlp-main.398/>
- [7] G. Sogancioglu, P. Mosteiro, A. A. Salah, F. Scheepers, and H. Kaya, "Fairness in ai-based mental health: Clinician perspectives and bias mitigation," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, no. 1, pp. 1390–1400, Oct. 2024. [Online]. Available: <https://ojs.aaai.org/index.php/AIES/article/view/31732>
- [8] S. Balloccu, P. Schmidová, M. Lango, and O. Dusek, "Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 67–93.
- [9] S. Balloccu, E. Reiter, K. J.-H. Li, R. Sargsyan, V. Kumar, D. Reforgiato, D. Riboni, and O. Dusek, "Ask the experts: sourcing a high-quality nutrition counseling dataset through human-AI collaboration," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 11 519–11 545. [Online]. Available: <https://aclanthology.org/2024.findings-emnlp.674/>
- [10] A. Welivita and P. Pu, "Curating a large-scale motivational interviewing dataset using peer support forums," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, Eds. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 3315–3330. [Online]. Available: <https://aclanthology.org/2022.coling-1.293/>
- [11] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, E. Reiter, D. Reforgiato Recupero, and D. Riboni, "Anno-mi: A dataset of expert-annotated counselling dialogues," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6177–6181.
- [12] Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, D. Reforgiato Recupero, and D. Riboni, "Creation, analysis and evaluation of annomi, a dataset

- of expert-annotated counselling dialogues,” *Future Internet*, vol. 15, no. 3, 2023. [Online]. Available: <https://www.mdpi.com/1999-5903/15/3/110>
- [13] V. Kumar, G. Medda, D. R. Recupero, D. Riboni, R. Helaoui, and G. Fenu, “How do you feel? information retrieval in psychotherapy and fair ranking assessment,” in *Advances in Bias and Fairness in Information Retrieval*, L. Boratto, S. Faralli, M. Marras, and G. Stilo, Eds. Cham: Springer Nature Switzerland, 2023, pp. 119–133.
- [14] X. Sun, J. Pei, J. d. Wit, M. Aliannejadi, E. Krahmer, J. T. Dobber, and J. A. Bosch, “Eliciting motivational interviewing skill codes in psychotherapy with LLMs: A bilingual dataset and analytical study,” in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 5609–5621. [Online]. Available: <https://aclanthology.org/2024.lrec-main.498/>
- [15] D. J. Min, V. Pérez-Rosas, K. Resnicow, and R. Mihalcea, “PAIR: Prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 148–158. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.11/>
- [16] D. J. Min, V. Perez-Rosas, K. Resnicow, and R. Mihalcea, “VERVE: Template-based ReflectIVE rewriting for MotiVational IntErviewing,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10 289–10 302. [Online]. Available: <https://aclanthology.org/2023.findings-emnlp.690/>
- [17] V. Kumar., S. Balloccu., Z. Wu., E. Reiter., R. Helaoui., D. Recupero., and D. Riboni., “Data augmentation for reliability and fairness in counselling quality classification,” in *Proceedings of the 1st Workshop on Scarce Data in Artificial Intelligence for Healthcare - SDAIH*, INSTICC. SciTePress, 2023, pp. 23–28.
- [18] V. Kumar, E. Ntoutsis, P. S. Rajawat, G. Medda, and D. R. Recupero, “Unlocking LLMs: Addressing scarce data and bias challenges in mental health,” 2024. [Online]. Available: <https://arxiv.org/abs/2412.12981>
- [19] V. Kumar, P. S. Rajwat, G. Medda, E. Ntoutsis, and D. R. Recupero, “Unlocking LLMs: Addressing scarce data and bias challenges in mental health and therapeutic counselling,” in *Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security*, R. Mitkov, S. Ezzini, T. Ranasinghe, I. Ezeani, N. Khallaf, C. Acarturk, M. Bradbury, M. El-Haj, and P. Rayson, Eds. Lancaster, UK: International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security, Jul. 2024, pp. 238–251. [Online]. Available: <https://aclanthology.org/2024.nlpaics-1.26/>
- [20] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *biometrics*, pp. 159–174, 1977.
- [21] M. Tanana, K. A. Hallgren, Z. E. Imel, D. C. Atkins, and V. Srikumar, “A comparison of natural language processing methods for automated coding of motivational interviewing,” *Journal of Substance Abuse Treatment*, vol. 65, pp. 43–50, 2016, motivational Interviewing in Substance Use Treatment. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0740547216000222>
- [22] K. A. Hallgren, “Computing inter-rater reliability for observational data: an overview and tutorial,” *Tutorials in quantitative methods for psychology*, vol. 8, no. 1, p. 23, 2012.