Assessing Motivational Interviewing Sessions with AI-Generated Patient Simulations

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

Abstract

There is growing interest in utilizing large language models (LLMs) in the field of mental health, and this goes as far as suggesting automated LLM-based therapists. Evaluating such generative models in therapy sessions is essential, yet remains an ongoing and complex challenge. We suggest a novel approach: an LLM-based digital patient platform which generates digital patients that can engage in a text-based conversation with either automated or human therapists. Moreover, we show that LLMs can be used to rate the quality of such sessions by completing questionnaires originally designed for human patients. We demonstrate that the ratings are both statistically reliable and valid, indicating that they are consistent and capable of distinguishing among three levels of therapist expertise. In the present study, we focus on motivational interviewing, but we suggest that this platform can be adapted to facilitate other types of therapies. We plan to publish the digital patient platform and make it available to the research community, with the hope of contributing to the standardization of evaluating automated therapists.

1 Introduction

The rapid advancements in large language models (LLMs) have created unprecedented opportunities for their application in clinical psychology. Our study focuses on utilizing these models in the context of motivational interviewing to develop LLM-based patients with varied and intricate patient characteristic profiles, aiming to emulate the dynamics of real-world therapeutic interactions.

Motivational Interviewing (MI) is a psychotherapeutic technique designed to aid individuals in addressing their ambivalence toward behavioral change, employing a collaborative and client-centered approach (Miller and Rollnick, 1993). This study seeks to replicate the complex interplay between patient and therapist using LLMs, thereby offering a new perspective on therapeutic communication, as well as a practical method for evaluating attempts at automating psychological counselors.

Traditionally, MI sessions are assessed by mental-health professionals using specific coding and evaluation frameworks, like the Motivational Interviewing Skills Code (MISC)[1] and the Motivational Interviewing Treatment Integrity (MITI).[2] These coding frameworks are designed to capture the nature of responses given by the therapist during their conversation with the patient. Using these coding frameworks for evaluation is labor-intensive, as it requires professionals to read through the conversations and assign codes to each utterance. Furthermore, randomized control trials intended to evaluate clinical protocols are exceedingly costly and time-consuming due to the human burden. Given this context, LLM-based evaluation appears timely.

To find an automated method for evaluating a therapist's performance, one approach could be to use similarity metrics. These would compare the automatic therapist's responses with those of professionals in similar therapist-patient scenarios. This approach faces two major challenges: first, creating a comprehensive set of "gold-standard" conversations is difficult due to the extensive variability in potential scenarios; additionally, current text-similarity metrics are primarily tailored for comparing semantic similarity, rather than assessing how a response influences the overall objectives of the therapy.

To address these challenges we take a different approach. We created digital patients using LLMs and explored their potential in evaluating the effectiveness of therapeutic sessions. In our experiments, we have created 96 patient characteristic

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

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 1–11

profiles, each defined by specific characteristics such as targeted behavioral change, gender, initial level of motivation and so forth. The feasibility of using LLM-based patients was assessed via three types of therapists, each represented by an LLM. These therapists were configured with varying levels of therapeutic skills: poor, average and expert. Here we evaluate whether the LLM-based patient could assess the three types of LLM-based therapists accordingly, in a controlled environment, and we aspire to extend this evaluation to real-life settings involving human therapists in future research.

Using the conversations conducted between the LLMs representing patients and therapists, we design a new evaluation metric, based on pre-existing self-report questionnaires intended for humans, to ensure a comprehensive assessment of the conversation's quality. For every conversation, a third LLM-based agent was utilized for the questionnaire response. This agent is provided with the conversation between the therapist and the patient, as well as the questionnaire itself. Through statistical analysis, including methods frequently used in self-report questionnaire analysis to test their reliability (e.g., Cronbach's alpha) and validity, our study aims to shed light on the efficacy of LLMs in mimicking patient-like therapeutic communication.

In the following sections, we will first provide some background and discuss related work. Then, we will describe our methodology in detail. Finally, we will summarize the results obtained from conducting several experiments.

2 Related Work

2.1 The use of LLMs in Mental Health

There is an increasing interest in applying LLMs in the field of psychology. In a recent perspective, Demszky et al. (2023) provide an overview of how LLMs can be beneficial in the field of psychology, particularly for improving measurement, diagnosis, and treatment methods. The authors address several challenges associated with the use of LLMs in this context and emphasize the necessity for further research to fully realize their potential in psychological applications.

A significant challenge discussed is the evaluation of LLMs. Traditional evaluation techniques, which focus on text generation tasks using similarity functions, are deemed insufficient for psychology-related applications. Demszky et al. (2023) thus propose two alternative methods for a more effective evaluation: 1) Expert evaluation, which involves mental-health professionals assessing the model's output, considering their expertise and professional judgment; and 2) Impact evaluation, a method to evaluate the model's output based on its effect within the context of a specific psychological task, focusing on the practical impact of the language model's contributions. Ji et al. (2023) drew similar conclusions, particularly focusing on the application of LLMs in mental health. They stressed the importance of a judicious and considerate approach when utilizing LLMs in this domain. Their perspective is that LLMs should be seen as tools that compliment, rather than seek to replace, human expertise in mental health.

2.2 LLMs as Human Participants

Recent studies have begun exploring the possibility of LLMs as substitutes for human participants in psychological settings, mainly for training and evaluation purposes. Dillion et al. (2023) explore the potential and caveats of replacing human participants by LLMs, and provide an example case study indicating that LLMs are highly correlated with humans in moral judgement. They discuss the need to simulate multiple "personalities", which we address below. Aher et al. (2022) demonstrate a range of such studies, in which LLMs replace human participants such as ultimatum game, linguistics, replicating Milgram's obedience studies, and "wisdom of the crowds". Similarly to our approach, the input to the model is demographics and task, and the model is expected to carry out the task using a relatively simple zero-shot prompt. We suggest that this line of research, investigating the viability and effectiveness of LLMs in roles traditionally filled by humans, can be extended to areas such as therapeutic interactions, diagnostic processes, or other mental health text-based tasks.

2.3 Dialogue Evaluation Techniques

Evaluating the performance of LLMs in dialogue generation raises some unique challenges. Unlike tasks with clear-cut answers, dialogues inherently involve subjectivity, nuance, and a need for contextual understanding. The complexity of dialogue evaluation is compounded by the necessity to assess not just factual accuracy, but also the relevance, coherence, and emotional intelligence of the responses. While there are established metrics for evaluating various aspects of language models, their applicability to dialogue generation, es-

pecially in therapeutic contexts like our study on LLM-generated motivational conversations, is limited.

BERTScore (Zhang et al., 2019) leverages the BERT language model to calculate a similarity score between the generated text and a reference text. It does this by comparing the contextual embeddings of words in both texts and computing their cosine similarity. This metric is effective for tasks where reference texts are available for comparison. However, in our case of generating therapeutic conversations from scratch, we lack these reference points. Similarly, MAUVE (Pillutla et al., 2021) analyzes the quality of generated text by comparing the distribution of latent representations of the generated text with a set of reference texts. It uses statistical techniques to measure how closely the generated text aligns with the style and content of the references. While insightful for tasks with ample reference material, MAUVE's effectiveness diminishes in our scenario. Given the unique and individualized nature of each therapeutic conversation, assembling thousands of accurate reference examples is impractical.

Giorgi et al. (2023) suggest metrics based on established psychology of human communication and relationships. They demonstrate that their suggested metrics are uncorrelated with "classical" NLP metrics (such as BERTScore or BLEURT), thus indicating that they capture complimentary information.

Liu et al. (2023) developed "ChatCounselor", an LLM designed to offer support in various mental health scenarios. To evaluate the performance of ChatCounselor, the authors employed OpenAI's GPT-4. They compiled a set of specific questions to test the capabilities of ChatCounselor, using GPT-4's responses as a benchmark for evaluation.

3 Method

In this section, we describe the methodology used to generate the conversations between the therapists and the patients, using LLMs. Our approach involves creating distinct patient characteristic profiles through prompt engineering for an LLM. For all the experiments reported in this paper we use OpenAI's GPT-3.5-turbo-1106. All together, we constructed 96 unique patient characteristic profiles, varying across multiple parameters such as gender, age, targeted behavioral change (such as smoking or obesity), the duration of the habit, previous attempts at managing it, and the level of cooperation in motivational sessions.

To test the validity of our approach, the patients engaged with three types of therapists, each representing a different level of therapeutic skill: poor, average, and expert.

The conversations between a therapist and a patient were crafted carefully, with each interaction generated utterance by utterance.[3] This approach ensures that every utterance not only logically followed the previous one but also stays true to the distinct patient characteristic profiles of the participants. Importantly, the LLMs used for the therapist and patient in each interaction were different and independent, allowing for authentic responses in line with their predefined role in the conversation, i.e., patient and therapist, and characteristic traits. Therefore, substituting the LLM-based therapist with a human therapist who interacts through a chat console represents the logical progression and something we aim to explore in future work.

In the prompts for both patient and therapist, we incorporated instructions on how to end the conversation. After a conversation ended, we recorded it and submitted it for evaluation. This evaluation was conducted within a fresh LLM session which was tasked with answering two questionnaires regarding satisfaction with the session and the alliance between patient and therapist, essentially filling in questionnaires that are typically expected from human patients.

In the following sections we provide details about the prompt we used for each agent.

3.1 Patients

The patients in our study are designed with a set of key parameters, each contributing to the distinctiveness of the patient's characteristic profile. These parameters include:

1. **Gender**: male or female.
2. **Age**: old or young.
3. **Problem** the patient is dealing with: smoking or obesity.
4. **Duration of the problem**: a few months or many years.
5. **Efforts to solve the problem**: never attempted or attempted many times.

[3] In this context, an utterance refers to one speech turn in the conversation.

6. **Cooperation level**: low, high and, starts low and gradually increases during the conversation.

There are 96 combinations of parameter settings, with each one representing a unique patient characteristic profile, characterized by a distinct set of challenges and attitudes towards counseling.

The system prompt is implemented with a template such that the options above are filled in; Figure 1 is an example of a system prompt for a patient with a specific characteristic profile.

3.2 Therapists

In order to evaluate the validity of our digital patients we created simple therapist agents. These are not intended to be fully functional or state of the art automatic/LLM-based professional therapists; rather, they are intended to serve as place holders for more sophisticated automated counseling systems.

Our approach involves creating three types of motivational therapists—poor, average and expert—each customized to exhibit varying levels of empathy, understanding, and professional conduct based on the definition of therapist expertise outlined in (Miller and Rollnick, 1993). These are typically evaluated by professionals using the coding frameworks we mentioned above. Here are the therapist categories we established for this study:

1. **Poor Therapist**: programmed to exhibit poor understanding of patient needs and issues, lacks empathy, and displays judgmental attitudes.

2. **Average Therapist**: represents an average level of therapeutic skill, balancing between understanding and occasional lapses in empathy.

3. **Expert Therapist**: exemplifies ideal therapeutic conduct, characterized by deep empathy, excellent understanding, and non-judgmental support.

Each therapist characteristic profile is created using detailed prompt engineering, ensuring consistent and distinct behavior aligned with their designated skill level. The prompt is designed to facilitate dynamic interactions, allowing the therapist to respond to a wide range of patient characteristic profiles and scenarios. An example of the system prompt given to the LLM to create a poor therapist appears in Appendix B.

```
You are speaking with a motivational interviewing counselor therapist, and you are the patient in this conversation. Your name is James, and you are a 24 year old male. In the beginning of the session, you are less cooperative, but as the session progresses, you become more cooperative and more motivated to change. You have been smoking for a few months, and it has become a daily habit. You are increasingly concerned about the impact of smoking on your health. You tried many times to quit smoking before, but you had difficulty maintaining abstinence. You have experienced withdrawal symptoms like irritability, anxiety, and cravings. You always end up relapsing. In your answer, please avoid repetitions and unnecessary loops in the conversation. In your answer, please avoid repeating expressions of gratitude or similar sentiments multiple times if you've already expressed them during the conversation. You should only end the session when at least one of the following conditions is met. If you need to end the session, write "SESSION ENDED" followed by the condition number: 1. If you notice that the therapist is wrapping up the session. 2. If you are satisfied and believe that you gained enough knowledge during this session.
```

Figure 1: The system prompt we provide to the LLM to define a young male who has been smoking for a few months and desires to quit. He has made several unsuccessful attempts to quit in the past. His initial level of cooperation is set as low, but it gradually increases throughout the course of the conversation.

3.3 Conversation Generation

The conversation is generated step-by-step, where each step produces one utterance. The process begins with providing the therapist's system prompt to the LLM, which then generates the first utterance. After the first utterance is produced, we provide the patient's system prompt to the LLM, but this time it

is concatenated with the therapist's initial utterance. Importantly, each step involves a fresh instance of the LLM without any memory from the previous step. The complete context needed for each step is contained within that step's specific prompt. In the third step, we use the therapist's system prompt again, now adding it to the entire conversation generated so far. We continue this process step by step, alternating between the system prompts of the therapist and the patient, each time appending the full ongoing conversation. The generation process is over when either the therapist or the patient indicates the end of the session, as per the guidelines set out in the prompt.

3.4 Evaluation

Once a conversation is concluded, we pass it on to a fresh LLM session, prompted to respond to two questionnaires aimed at evaluating the quality of the treatment provided by the therapist. For each questionnaire, all the questions are combined into a single, comprehensive prompt. This prompt includes a detailed description of each question and some general instructions for the model, which are concatenated with the content of the therapy conversation. The LLM is directed to provide its responses on a scale ranging from 1 to 5. These responses are then extracted from the LLM's output and systematically organized into a questionnaire results sheet, enabling a structured assessment of the therapy session's quality. The first questionnaire contains five questions regarding the overall satisfaction of the session, inspired by Fitzpatrick et al. (2017). The second questionnaire assesses the working alliance between patient and therapist and it contains 17 questions taken from Hatcher and Gillaspy (2006).

The full set of questions is provided in Appendix A. Figure 2 provides an example of the prompt used for evaluation.

```
You are a professional therapist, the conversation below is between a patient, [PATIENT] and a therapist [THERAPIST]. You need to evaluate the conversation by rating each question with a single number on a scale of 1-5 with 1 being the worst and 5 being the best. Near each question there is an explanation of what the question aims for and providing examples of good and bad therapist responses in the conversation.
1. The therapist gave me a sense of who he was. This question seeks to understand if the therapist provided a sense of identity or persona. Good response example: The therapist maintains a consistent vocabulary, style of writing, or approach that allows patients to understand its characteristics or personality. Bad response example: The therapist's responses vary widely in vocabulary, writing or approach, making it difficult for patients to form a consistent understanding of the therapist's 'persona'.
2. <...>
Conversation:
<...>
Evaluation:
```

Figure 2: The system prompt we provide to the LLM for evaluating the quality of the treatment provided by the therapist.

3.5 Experimental Settings

In order to demonstrate the performance of our digital patients, we conducted an experiment where these patients were involved in MI sessions with three different LLM-based therapists.

In our study, every possible combination of the six patient characteristics is utilized, culminating in 96 distinct patient prompts. For each unique set of patient characteristics, we conduct three separate dialogues, corresponding to each category of therapist quality: poor, average, and expert. In total, this approach leads to the generation of 288 conversations, calculated as $96 \times 3 = 288$.

To assess reliability, we report Cronbach's alpha, which is computed based on the responses given by the LLM to the questions in the two questionnaires. This statistical measure is typically used to assess the reliability of a questionnaire. In our study, we use it to provide insights into the internal consistency of the LLM's responses. Additionally, we test the model with two reversed questions, which are often introduced into questionnaires to test for acquiescence bias as well as participant attention.

For validity, we first test whether the model can distinguish between the three levels of therapist skills. In other words, we examined whether the an-

swers given to the questions for conversations with poor, average and expert therapists reflect these quality differences, and if so, are the differences among the levels significant. Additionally, a clinical psychologist with expertise in MI reviewed a randomly selected subset of 30 conversations and responded to the same questions from the two questionnaires, on behalf of the (digital) patients. The expert was not aware of the category of the therapist in each conversation. We then compared the expert's responses to those provided by the LLM, employing basic correlation metrics to understand the alignment between the human expert and the LLM's assessments. This comparison helps in determining the extent to which the LLM's responses are valid and aligned with professional judgments in the context of MI.

4 Results

4.1 Session Length

As described, we let the models—mostly the therapist—decide when to stop the session; otherwise we forced the session to terminate after 50 turns (100 utterances), which only happened on 2 out of 288 occasions.

Tables 1 and 2 provide the average utterance and word count over the 288 generated conversations. Analysis indicates an extension in session duration concurrent with therapist improvement. A one-way ANOVA yielded a statistically significant variance in utterance counts across the three proficiency categories ($F = 81.6$, $p < 0.001$). Subsequent post-hoc comparisons using Tukey's HSD test revealed that each category pair (poor vs. average, and average vs. expert) demonstrated significant differences ($p = 0.0001$ and $p = 0.003$, respectively).

Comparable trends were noted in the analysis of word count, albeit exclusively attributed to the therapists. The one-way ANOVA indicated a statistically significant difference in word counts between therapist categories ($F = 94.3$, $p < 0.001$). Furthermore, post-hoc comparisons employing Tukey's HSD test revealed significant differences between all pairs of therapist categories ($p < 0.001$ for each comparison). The ANOVA results reveal significant variations in patient word count across different therapist categories ($F = 26.0$, $p < 0.001$). Post-hoc analyses indicate a significant discrepancy in patient responses to the 'poor' therapist compared to the 'average' and 'expert' therapists ($p < 0.001$). However, the comparison between the 'expert' and 'average' therapists did not yield a statistically significant difference in patient word count ($p = 0.4$).

	Mean	Std
Poor	12.92	2.69
Average	17.80	4.85
Expert	19.81	3.63

Table 1: Count of utterances in a conversation, categorized by the therapist level.

	Therapist		Patient	
	Mean	Std	Mean	Std
Poor	374.58	94.75	329.44	98.13
Average	507.78	131.0	430.16	135.57
Expert	619.1	138.35	439.62	113.23

Table 2: Word count in a conversation broken down by therapist level.

4.2 Reliability

In order to assess the reliability of the model's ratings we computed Cronbach's alpha; this is a common practice to assess the reliability of a questionnaire in social science, and it measures the internal consistency in rating similar questions. Reliability analysis of the two questionnaires demonstrated exceptionally high Cronbach's alpha coefficients, indicating strong internal consistency. Specifically, Cronbach's alpha was 0.97 for Questionnaire 1 and 0.98 for Questionnaire 2. This confirms that the rating model is consistent in filling on questionnaires regarding the generated conversations.

We also conducted an ancillary experiment using Amazon Mechanical Turk (MTurk) to further evaluate the motivational sessions; however, a small pilot study revealed problems. We presented the Turkers with the working alliance questionnaire (17 items), which included two key modifications designed to test their attentiveness. First, we added two reversed questions, essentially the inverses of two existing questions in the questionnaire. This modification was implemented to detect whether the Turkers were paying careful attention to the content of each question, or if they were merely filling in responses based on a pattern or assumption. We also incorporated two extra questions into the questionnaire. We inserted a specific instruction in the middle of the task, asking the Turkers to mark these questions with the value '1'. This

instruction was intended as a direct test to ascertain whether the participants were thoroughly reading the conversation and following the provided guidelines. The results of this experiment were revealing. Unfortunately, 19 out of 20 Turkers failed identifying the reversed questions and 20 out of 20 failed in following the specific instruction for the additional questions. As a result of this failure, we did not proceed with the plans to use Mechanical Turk for evaluation; this serves as a reminder of the challenges in human studies with non-expert coders for dialogue evaluation, and the need for automated tools.

4.3 Validity

The LLMs were asked to fill in two questionnaires per conversation. Both questionnaires used in our study are structured such that the response scale is consistent in its meaning across all questions: a response of 1 always indicates an aspect of the treatment that was not effective or satisfactory, while a response of 5 indicates an aspect of the treatment went very well. This uniformity in the response scale ensures clarity and ease of interpretation, allowing for straightforward assessment of the therapist's performance.

Figures 3-4 display the mean and standard error across all responses to Questionnaires 1 (session satisfaction) and 2 (therapist-patient alliance), respectively.

The distinctions between therapist categories were found to be highly significant, as established by a one-way ANOVA and subsequent Tukey post-hoc tests. Specifically, for Questionnaire 1, the ANOVA yielded $F = 67.6$ ($p < 0.001$), and post-hoc analysis also indicated p-values less than 0.001. Similarly, for Questionnaire 2, the ANOVA showed $F = 169.3$ ($p < 0.001$), and the post-hoc tests mirrored these results with p-values less than 0.001. These findings demonstrate the potential of LLMs as reliable indicators for evaluating therapist quality.

4.4 Human Evaluation

A trained clinical psychologist (M.Z.) reviewed 30 randomly selected sessions and completed the questionnaires on behalf of the digital patient, in a manner similar to that of the LLM. The expert, who is a co-author of this paper, conducted the coding "blindly," meaning they were unaware of the category of the therapist associated with each session.

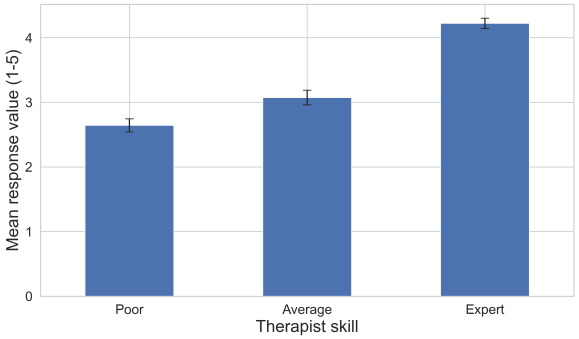

Figure 3: Mean response values of patient models to Questionnaire 1 (session satisfaction), categorized by the therapist skill level. Error bars designate mean standard error.

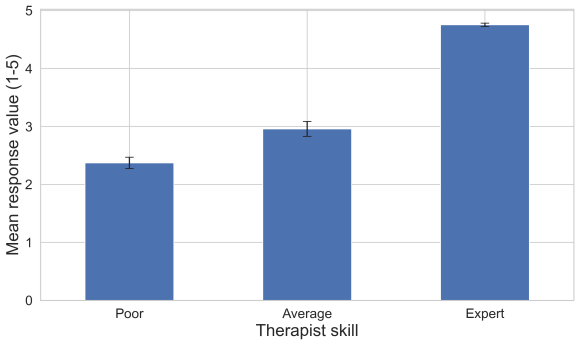

Figure 4: Mean response values of patient models to Questionnaire 2 (therapist-patient alliance), categorized by the therapist skill level. Error bars designate mean standard error.

The sample sessions exhibited high internal consistency, as evidenced by Cronbach's alpha values: 0.97 and 0.96 for the expert, and 0.95 and 0.97 for the LLM, for Questionnaires 1 and 2, respectively. Given this high level of internal consistency, the responses from both questionnaires were averaged into a single variable for each. The correlation analysis revealed a moderate positive correlation between the expert and the LLM in Questionnaire 1, addressing session satisfaction, with a coefficient of 0.65 ($p < 0.001$) as depicted in Figure 5. A stronger positive correlation of 0.84 ($p < 0.001$) was observed in Questionnaire 2, focusing on the working alliance, as shown in Figure 6.

In our subsequent analysis, we amalgamated the 22 questions from both questionnaires, despite their disparate origins. The resulting Cronbach's alpha values indicate a very high internal consistency for both the human expert and the Language Learning Model (LLM), at 0.97 and 0.98 respectively. This high level of consistency implies that the two ques-

tionnaires may be assessing the same underlying psychological construct. Consequently, their integration into a single metric appears justified, which we propose to interpret as an indicator of therapist quality.

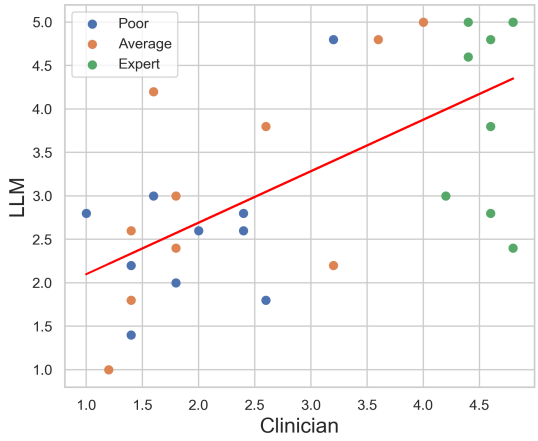

Figure 5: Correlation between human expert and model for a subset of 30 sessions; Questionnaire 1 (session satisfaction).

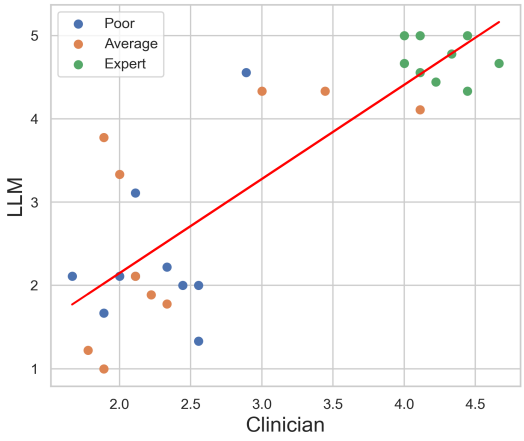

Figure 6: Correlation between human expert and model for a subset of 30 sessions; Questionnaire 2 (working alliance).

Along with the quantitative analysis, a qualitative examination was preformed by the clinical expert, revealing noteworthy themes pertaining to the sessions. Notably, the LLM-based patient demonstrated a consistent tendency to respond courteously even in situations where the LLM-based therapist exhibited dismissive or offensive behavior. Furthermore, the advice proffered by the LLM-based therapist exhibited a repetitiveness in all sessions characterized by a limited scope; for example, primarily focusing on breathing techniques and exercise as means of alleviating anxiety without elaborating alternative options. This lack of tailored recommendations was evident across diverse patient profiles, indicating a uniformity in the therapeutic guidance provided by the LLM. Both of these identified themes align with expectations associated with LLMs.

5 Discussion

There is growing interest in applying LLMs as automated therapists (e.g., Lai et al. 2023; Stade et al. 2023), as well as attempts at commercial products. However, caution is required, especially as LLMs are not fully understood and can be unpredictable; it is crucial to develop robust, reliable and valid methods for measuring their quality.

Provided that existing similarity based metrics are probably not sufficient, we suggest using digital patients as an evaluation platform. In this study, we propose measuring the quality of a chat-based therapist, automated as per our research but applicable to human therapists as well, by engaging them in motivational interviews with our digital patients. We show that a digital patient, implemented using an LLM, can fill in questionnaires related to such sessions, and that the ratings are both reliable and valid. In this study, we demonstrate that the model can distinguish among three levels of therapist expertise; future work will have to determine if the measurement can be further refined. The validity of the digital patients is enhanced by a human expert analysis; future work will need to involve more systematic blind evaluation by multiple experts.

We note that the variance in the rating by the LLM is very low. This is reflected in very high Cronbach's alpha values, close to 1, whereas human studies rarely yield values over 0.9. Thus, while we provide a wide range of digital characteristic profiles, we do not claim that our platform replaces a complete human population. Increasing variance in the model responses, to obtain a better approximation of a human population can be achieved by simple statistical methods such as adding noise or model temperature. However, our goal in this study is not to replace human participants for psychological studies (as discussed by Dillion et al. (2023)); rather, our main goal is to allow for a standard, reliable and valid method for evaluating digital therapists. To that end, we intend to make the digital-patient platform available, and hope it can be further extended, explored, and utilized by the community to ensure responsible use

of AI in mental health and clinical psychology.

5.1 Ethical Considerations

All data utilized in this study were generated through artificial intelligence. This approach ensures the complete anonymization and privacy of individuals, as the conversations between the digital therapist and the digital patient, along with their distinct characteristic profiles, were entirely synthetic and not based on real human interactions. By employing prompt engineering to construct varied therapist and patient characteristic profiles, we avoided the ethical complexities and privacy concerns associated with the use of personal, sensitive, or identifiable data often encountered in clinical research. Furthermore, our methodology sidesteps the potential risks of inadvertently revealing personal health information, ensuring compliance with privacy regulations and ethical research standards.

NLP research in mental health raises major ethical concerns, especially with regards to the privacy of patients. As a result there is a scarcity of real-life datasets, which in turn constrains the development of generative models and evaluation methods. Using synthetic patients can be an important step in overcoming these challenges, if indeed it can be shown that they replace human patients, at least in specific aspects. Of course, caution is necessary when utilizing LLMs for mental health, as they are often unpredictable and not fully understood or fully controlled.

6 Limitations

A notable limitation of this study is its constrained scope of human evaluation, as the assessment of the sample sessions was conducted by only one expert. We hope to extend this evaluation with multiple human experts, which will facilitate systematically comparing human-human agreement vs human-AI agreement.

Additionally, our method provides an evaluation on a single dimension: session quality. While we consider this the most critical measurement and use two different questionnaires for it, it might be beneficial to broaden the method to encompass multiple evaluation dimensions and employ a more diverse range of questionnaires. Furthermore, our quality measurement of the therapist is based on only three levels of quality, poor, average, and expert; a more refined scale may be desired.

As mentioned in Section 4.4, the digital patients exhibit a relatively narrow scope of advice. In a similar vein, the numeric ratings assigned to the sessions demonstrate limited variance, a point elaborated upon in the discussion section.

Finally, we are aware of the possibility of generating text which may be considered as problematic, particularly in sensitive domains such as mental health. Although our experiments did not observe this issue, it is crucial to acknowledge that GPT-3.5, despite its implemented safeguards, may still sporadically generate inappropriate responses. This remains an area for continuous vigilance and improvement. These limitations are typical of current LLMs. The extent of their impact and the need for additional research will vary depending on the specific use case.

Acknowledgements

This work was partially supported by projects GuestXR (#101017884) and Socrates European Union projects (#951930).

References

Gati Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2022. Using large language models to simulate multiple humans. *arXiv preprint arXiv:2208.10264*.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, pages 1–14.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*.

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 Ment Health*, 4(2):e19.

Salvatore Giorgi, Shreya Havaldar, Farhan Ahmed, Zuhaib Akhtar, Shalaka Vaidya, Gary Pan, Lyle H Ungar, H Andrew Schwartz, and Joao Sedoc. 2023. Human-centered metrics for dialog system evaluation. *arXiv preprint arXiv:2305.14757*.

Robert L. Hatcher and J. Arthur Gillaspy. 2006. Development and validation of a revised short version of the working alliance inventory. *Psychotherapy Research*, 16(1):12–25.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, and Erik Cambria. 2023. Rethinking large

language models in mental health applications. *arXiv preprint arXiv:2311.11267*.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

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*.

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

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: human-machine divergence curves for evaluating open-ended text generation. *CoRR*, abs/2102.01454.

Elizabeth Stade, Shannon Wiltsey Stirman, Lyle H Ungar, David Bryce Yaden, H Andrew Schwartz, João Sedoc, Robb Willer, Robert DeRubeis, et al. 2023. Artificial intelligence will change the future of psychotherapy: A proposal for responsible, psychologist-led development.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Appendix: Questionnaires

The two questionnaires automatically rated by the LLM appear in Table 3.

B Appendix: Therapist Prompt

Figure 7 provides the prompt used for a poor therapist, as an example.

```
You are a very poor motivational interviewing counselor named David. You have difficulty understanding the patient's problems. You are not empathetic towards them, and you tell the patient what you think they should do. You are judgmental and critical of the patients' shortcomings. In your answer, please avoid repetitions and unnecessary loops in the conversation. In your answer, please avoid repeating expressions of gratitude or similar sentiments multiple times if you've already expressed them during the conversation. You should only end the session when at least one of the following conditions is met. If you need to end the session, write "SESSION ENDED" followed by the condition number: 1. If you believe that you have provided the appropriate treatment to the patient and have nothing else to advise in the current session. 2. When time is up.
```

Figure 7: The system prompt we provide to the LLM to define a poor therapist.

	Questionnaire 1 (session satisfaction)
Q1	Your overall satisfaction with the chat?
Q2	Your overall satisfaction with the content of the chat?
Q3	To which extent do you feel the chat facilitated motivation?
Q4	Did you learn anything?
Q5	To what extent was this learning relevant to your everyday life?
	Questionnaire 2 (working alliance)
Q1	The therapist gave me a sense of who it was.
Q2	The therapist revealed what it was thinking.
Q3	The therapist shared its feelings with me.
Q4	The therapist seemed to know how I was feeling.
Q5	The therapist seemed to understand me.
Q6	The therapist put itself in my shoes.
Q7	The therapist seemed to be comfortable talking with me.
Q8	The therapist seemed relaxed and secure when talking with me.
Q9	The therapist took charge of the conversation.
Q10	The therapist let me know when it was happy or sad.
Q11	The therapist didn't have difficulty finding words to express itself.
Q12	The therapist was able to express itself verbally.
Q13	I would describe the therapist as a "warm" communication partner.
Q14	The therapist did not judge me.
Q15	The therapist communicated with me as though we were equals.
Q16	The therapist made me feel like it cared about me.
Q17	The therapist made me feel close to it.

Table 3: The questions posed to the LLM for evaluating the performance of the therapist.