

LLMs as Standardised Patients for Motivational Interviewing: How Faithful Are They?

Van Hoang^{1,2} and Eoin Rogers¹ and Robert Ross¹

¹School of Computer Science, Technological University Dublin

²ML-Labs, SFI Centre for Research Training in Machine Learning, Ireland

van.hoang, eoin.rogers, robert.ross@tudublin.ie

Abstract

Recent advances in large language models (LLMs) have enabled the creation of highly realistic digital patients across a broad range of clinical scenarios, yet systematic evaluation of such simulations remains challenging due to a lack of standardised methodology. This paper investigates the faithfulness of LLM-simulated patients within motivational interviewing contexts. We directly compare the properties of data generated by simulated and human patients given identical profiles, rather than relying on subjective user experiences. Our findings reveal that while simulated and human patients produce semantically similar content and engage with comparable topics, their modes of expression differ substantially. LLM-simulated patients struggle to reproduce the full complexity of human behaviours and attitudes. While human patients exhibit a mix of positive and negative responses, LLM patients skew toward uniformly positive ones.

1 Introduction

User simulation, the modelling of user preferences and behaviours from observational data, has long been an active area of research, with applications spanning a wide range of domains (Ni et al., 2025). In clinical AI, this approach is applied to create digital patients, which can be used to train healthcare providers and evaluate dialogue systems.

Traditionally, in medical education, standardised, or simulated, patients are individuals carefully trained to portray real patients, giving students the opportunity to develop and be evaluated on clinical skills in a realistic setting (Flanagan and Cummings, 2023). As a widely adopted method, it offers valid and reliable assessment of both clinical and communication competencies across all levels of healthcare training (Plaksin et al., 2016). However, their large-scale implementation is constrained by high costs and the psychological burden

placed on individuals repeatedly portraying mentally ill patients (Plaksin et al., 2016). Computer-based simulations have emerged as a scalable alternative. Recent advances in LLMs have enabled the development of highly realistic digital patients across diverse clinical scenarios, offering new opportunities for simulation-based medical education (Chu and Goodell, 2024; Zeng et al., 2025).

Additionally, researchers have advocated using simulated patients to evaluate conversational AI models for real-world applications (Wang et al., 2024a; Liao et al., 2024). Traditional dialogue systems are typically assessed by measuring utterance-level performance on held-out test sets. It is, however, fundamentally ill-suited for therapeutic contexts, where dialogue is inherently nuanced and demands contextual understanding. Unlike tasks with clear-cut answers, a single utterance can elicit a wide range of valid responses: experts have verified that even when predicted strategies diverge from ground-truth annotations, they can still align effectively with therapy principles and context (Sun et al., 2025). Simulated patients therefore offer a particularly promising framework for systematic, dialogue-level evaluation of conversational systems, enabling assessment that extends well beyond single-turn exchanges.

Despite their promise, the utility of LLM-based patient simulation hinges on whether simulated patients can faithfully and validly represent real human patients. Evaluating LLM-simulated patients has largely relied on user experience, whether from laypeople or domain experts, who converse with the simulation or review transcripts and rate dimensions such as satisfaction, coherence, realism, or learning outcomes (Ni et al., 2025; Zeng et al., 2025). Prior work suggests that LLM-generated utterances are natural, coherent (Galland et al., 2024), and perceived as human-like (Hämäläinen et al., 2023). Expert evaluations have further demonstrated LLMs' potential to authentically simulate

mental health clients (Wang et al., 2024b; Steenstra et al., 2025), though the field still lacks validated and standardised assessment tools (Zeng et al., 2025).

This work therefore aims to assess how reliably LLM-based patient simulations can represent real patients in the context of motivational interviewing-centric dialogue systems. Rather than relying on subjective experience ratings, we evaluate simulated patients by directly comparing the data they produce against that of real human patients to identify key differences and similarities. Specifically, we operate under the assumption that a simulated patient is a faithful representation of a human patient if, given the same case vignettes, the properties of the data each generates are comparable. Using a motivational interviewing (MI) coaching dataset, we extract user profiles to instantiate LLM-based simulated patients, which then interact with MI systems to generate simulated conversations for comparison against real human data.

MI is a psychotherapeutic technique designed to foster motivation for behavioural change in ambivalent patients (Miller and Rollnick, 2023). We choose MI due to its growing popularity in the AI counselling domain, which has produced multiple openly available resources, including MI dialogue datasets and systems (Meyer and Elswiler, 2024; Yang et al., 2025a; Hoang et al., 2026), patient simulation platforms (Yosef et al., 2025; Yang et al., 2025b), and analysis tools (Galland et al., 2025), some of which we utilise in our experiments.

Our contributions include a comparison of LLM-simulated and human patients across linguistic patterns, dialogue act distributions in MI sessions, and questionnaire responses. Our analysis reveals behavioural and attitudinal differences between the two, which we attribute to an agreeableness bias in LLMs. Finally, we discuss approaches to more accurately model patient states and guidances for the future design of digital patients.

2 Related Works

2.1 Motivational Interviewing

MI is an evidence-based therapeutic approach designed to reduce resistance and foster clients' intrinsic motivation for behavioural change (Miller and Rollnick, 2023). Central to MI is the skill to evoke "change talk", considered as the patients' verbal expressions of commitment to change, which has been identified as a key predictor of behavioural

change and therefore represents a crucial focus for practitioners (Amrhein et al., 2003).

Building on these principles, digital health coaches that leverage MI have emerged as a growing area of research, demonstrating effectiveness across a range of behavioural intervention domains (Shingleton and Palfai, 2016; Pedamallu et al., 2022; Mercado et al., 2023), including eating habits (Robinson et al., 2020), alcohol use (Steenstra et al., 2024), smoking cessation (Brown et al., 2023), and physical activity (Jörke et al., 2025).

2.2 LLM-Simulated Personas

User simulations leverage LLMs to embody specific personas or characters, enabling dynamic, interactive exchanges that approximate human behaviour. Strong correlations between LLM and human responses across a range of psychological tasks support the use of LLMs as proxies for human patients in psychological research (Dillion et al., 2023; Grossmann et al., 2023). Further validating this approach, multiple researchers have successfully replicated classic behavioral science findings using ChatGPT, suggesting the viability of LLMs as scalable and cost-effective research tools (Aher et al., 2023).

In the domain of AI therapy, LLM-based user simulation has been proposed for three broad purposes: synthetic data collection, medical education, and the evaluation of dialogue systems. Regarding data collection, Galland et al. (2024) employed LLM-simulated patients to generate synthetic MI data for training MI dialogue systems, overcoming the difficulty of obtaining real therapy data. In medical education, LLM-based patient simulation has shown promise for training counsellors across experience levels (Zeng et al., 2025). Wang et al. (2024b) developed PATIENT- ψ , a simulation platform that emulates authentic patient communication patterns for cognitive behavioural therapy training. Similarly, Steenstra et al. (2025) integrated an LLM with an MI patient cognitive model to simulate patients in real time, alongside an interface capable of delivering expert feedback through data visualisation tools.

Beyond training, researchers have advocated for simulated clients as an evaluation mechanism for conversational AI (Wang et al., 2024a; Liao et al., 2024), offering an alternative to conventional automated metrics such as BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2020). This approach has enabled Johri et al. (2025) to assess

clinical LLMs on diagnostic abilities within interactive conversations, and informed the development and validation of digital patients as an evaluation framework for chat-based therapy sessions (Yosef et al., 2025; Yang et al., 2025b).

3 Methodology

We now describe our human data source, patient simulation methodology, and methods for evaluating simulation faithfulness.

3.1 Data

3.1.1 Human Data

Our human data source comes from Hoang et al. (2026), a study in which participants interacted with two LLM-based MI coaches through a text interface on the topic of physical activity, followed by a questionnaire measuring their therapeutic alliance with the coach. Participants were recruited through Prolific and manually screened for quality, yielding a final dataset of 140 participants. From this dataset, we extract user profiles to create LLM patients, and have them interact with the same MI coaches described in the paper, generating simulated conversations for comparison against the real human data. Further details are provided in the following section.

3.1.2 Simulated Data

Creating MI Patient Profiles. We adapt the MI patient simulation framework proposed by Yosef et al. (2025) for our experiments. Their work demonstrated that simulated patients could reliably distinguish between varying levels of therapist expertise and complete therapy-related questionnaires, with ratings shown to be both reliable and valid. Each simulated patient is characterised by the following parameters:

1. **Gender:** male or female.
2. **Age:** 18–64 years old.
3. **Problem:** the problem the patient is facing.
4. **Efforts to solve the problem:** never attempted or attempted many times
5. **Cooperation level:** low, high, or starts low and increases during the conversation.
6. **Intensity activity level:** low, or moderate.

You are speaking with a physical activity coach, and you are the client in this conversation. You are a 26-year-old male. You have been experiencing knee pain, which has significantly decreased your physical activity levels. Despite numerous attempts to increase your exercise routine, a lack of motivation has held you back. You are seeking guidance to overcome these barriers and improve your fitness. Initially, you are less cooperative, but as the session progresses, your willingness to engage and find solutions increases.

In your answer, please avoid repetitions and unnecessary loops in the conversation. Reply as in a chat interface and keep your response brief and short around 5-15 words long, like chatting with a health provider. Avoid repeating expressions of gratitude or similar sentiments multiple times if you've already expressed them during the conversation. The output should be a string of the patient's response and nothing else. Do not generate any opening, closing, and explanations. You should only end the session when at least one of the following conditions is met. If you need to end the session, write "Good bye." 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 have a realistic plan for your behaviour change. 3. You hate to change and want to stop the session.

Table 1: Example of a patient simulation prompt.

Three parameters (i.e., “Gender”, “Age”, and “Intensity activity level”) are obtained directly from participant demographic data, while the remaining parameters are extracted from the conversational data by GPT-4o¹. Table 1 illustrates an example prompt to simulate an LLM patient.

Generating Simulated Conversations. After constructing patient profiles, we then prompt GPT-4o to simulate MI patient behaviours by having it interact with the MI coaches (Hoang et al., 2026). A dialogue $d = (p_0, c_0, p_1, \dots, p_i, c_i)$ is defined as an alternating sequence of patient turns p_i and coach turns c_i , where each turn's textual content is referred to as an utterance. At each turn i in the dialogue, the simulated patient utterance is generated by conditioning on the current dialogue history and

¹Version gpt-4o-2024-08-06

Dialogue Act	Definition
Changing unhealthy behaviour	The patient explicitly expresses their willingness to change
Sustaining unhealthy behaviour	The patient explicitly expresses their unwillingness to change
Sharing negative feeling or emotion	The patient shares a negative feeling or vision of the world
Sharing positive feeling or emotion	The patient shares a positive feeling or vision of the world
Sharing personal information	The patient shares factual personal information about their situation or background
Understanding or realisation	The patient realizes or understand something about their problem
Asking for information	The patient asks for further information
Greeting or closing	The patient opens or closes the conversation
Backchannel	The patient acknowledges that they heard the last coach’s statement

Table 2: MI Patient Dialogue Acts (Galland et al., 2025).

the corresponding patient profile. A dialogue concludes when (1) the coach initiates a wrap-up, (2) the patient confirms their change plan, or (3) the patient initiates the end of the conversation. We generate 140 conversations, which we will then compare with the actual participant interactions.

3.2 Simulated Patients Evaluation

For simulated patients to be faithful representations of real human patients, their data, given identical profiles, should exhibit similar behavioral properties. We evaluate this across three dimensions: (1) linguistic patterns, (2) distribution of dialogue acts, and (3) questionnaire responses. Our assessment is at the dialogue level rather than the turn level. While turn-by-turn comparison is possible, a single utterance can elicit a range of valid responses (Sun et al., 2025), meaning the order and timing of utterances need not be identical across simulations, provided their content remains consistent with the patient’s profile.

Linguistics. We evaluate semantic and stylistic similarity using SimCSE by Gao et al. (2021) and style embeddings by Wegmann et al. (2022). While semantic embeddings measure the content of texts (i.e., what is said), style embeddings capture linguistic style (i.e., how it is said). The style embeddings employed in this work are trained on an authorship verification task, which predicts whether two texts were written by the same author. Thus, even if two texts share similar content and topics, their stylistic similarity score can be low if their writing styles differ.

Dialogue Acts. Each patient utterance is clas-

sified into a dialogue act. We utilise a list of 9 MI-specific patient dialogue acts, presented in Table 2, using classification prompts rooted in prior work by Galland et al. (2025). These prompts were developed and validated against manual annotations of a subset of an MI dataset, achieving an average accuracy of 0.9 and F1 score of 0.7.

Using GPT-5.2, we extract predicted dialogue acts for both human and simulated conversations. Rather than performing turn-level comparisons between human and LLM-generated patient dialogue acts, we compare their distributions across entire dialogues. Our intuition is that while individual turn-level actions may differ, the overall dialogue-level distributions is a more meaningful measure of behavioral alignment.

Questionnaire Responses. We investigate whether simulated patients produce questionnaire responses comparable to those of human patients. After obtaining simulated conversations, we collect responses from LLM patients for the Working Alliance Inventory Short Form (WAI) (Hatcher and Gillaspy, 2006), which measures therapeutic relationship between coaches and patients. The WAI is a 12-item self-report instrument that employs a 5-point Likert scale, with responses ranging from 1 (seldom) to 5 (always). Higher scores indicate a stronger working alliance, which consistently predicts better outcomes across interventions.

In their study, Yosef et al. (2025) compared the WAI scores generated by the LLM against those scored by three MI expert evaluators, finding strong positive correlations using Pearson’s coefficient. Similarly, Li et al. (2024) collaborated with 9 psy-

Dialogue Acts	Simulated Patients	human patients
Changing unhealthy behaviour	47.91 (\pm 14.44)	23.99 (\pm 15.72)
Sustaining unhealthy behaviour	0.23 (\pm 1.25)	0.64 (\pm 2.07)
Sharing negative feeling or emotion	2.15 (\pm 3.94)	4.04 (\pm 5.33)
Sharing positive feeling or emotion	7.92 (\pm 6.59)	3.87 (\pm 5.40)
Sharing personal information	24.97 (\pm 11.19)	29.35 (\pm 12.40)
Understanding or realisation	3.53 (\pm 4.51)	1.95 (\pm 3.45)
Asking for information	2.95 (\pm 4.83)	13.18 (\pm 13.70)
Greeting or closing	7.57 (\pm 8.07)	9.78 (\pm 6.51)
Backchannel	2.76 (\pm 3.53)	13.19 (\pm 10.69)

Table 3: Mean and standard deviations of patient behaviours.

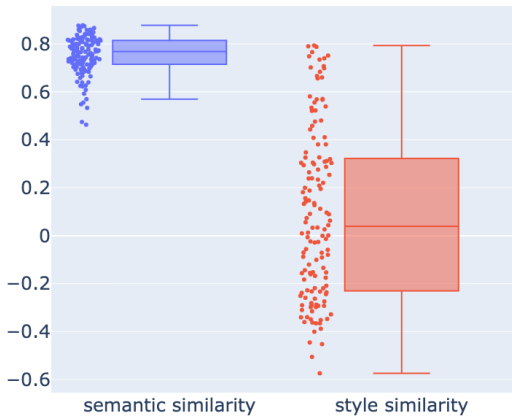


Figure 1: Semantic similarity and stylistic similarity scores for 140 conversation pairs of simulated patients and human patients.

chologists to develop guidelines and prompts for automatically assessing therapeutic alliance in text-based counselling sessions. We adopt the framework of Li et al. (2024), prompting GPT-5.2 to generate WAI scores automatically. The prompts were designed by expert psychologists, and the resulting scores achieve a Pearson correlation of approximately 0.5 against gold-standard data. Table 6 in Appendix A presents the template prompt for the WAI questionnaire, and Table 7 shows an example score rubric for Question 1. The complete rubrics for all WAI questions are provided in Appendix A.3 in Li et al. (2024)’s paper.

4 Results

4.1 Linguistics

Figure 1 illustrates the semantic and stylistic similarity scores at the dialogue level for 140 conversa-

tion pairs between simulated patients and human patients. Dialogue-level scores are computed by concatenating all utterances from LLM patients or human patients within each conversation. The figure shows that semantic similarity scores are consistently high (median \approx 0.75), with most conversation pairs clustered between 0.7 and 0.85, indicating that simulated patients and human patients discuss similar content. In contrast, stylistic similarity scores are substantially lower and more variable (median \approx 0.05), ranging from approximately -0.6 to 0.8, suggesting considerable variation in how conversations are expressed. The large gap between semantic and stylistic similarity demonstrates that while LLM patients can match the conversational content of human patients, their linguistic style differs greatly across individual conversations. Where each human brings a distinct writing style, all LLM patients share one.

4.2 Dialogue Acts

Table 3 presents the distributions of dialogue acts of both human and simulated MI patients. The most pronounced difference is in “*Changing unhealthy behaviour*”, which is nearly twice as frequent in simulated patients (47.91%) compared to human patients (23.99%), suggesting that simulated patients are overly positive in their expression. Conversely, human patients are more likely to ask for information (13.18% vs. 2.95%) and produce backchannels (13.19% vs. 2.76%), indicating that simulated patients are less interactive and reactive in conversation. This may reflect an inherent limitation of simulated patients: unlike human patients who sometimes seek clarification or ask follow-up questions when presented with medical information, simulated patients appear to process all information perfectly. Other dialogue acts, such as

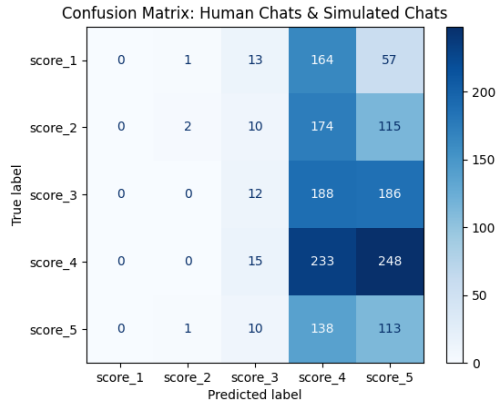


Figure 2: The confusion matrix of WAI scores between simulated patients and human patients.

sharing personal information and sharing positive or negative feelings, are more comparable across the two groups.

4.3 Questionnaire Responses

We employ the prompt described above to obtain WAI scores for simulated conversations and compare these scores against those from human patients. The Pearson correlation coefficient between simulated and human scores was 0.14, indicating weak alignment. Figure 2 presents the confusion matrix for 1680 WAI score pairs between simulated patients and human patients. The matrix reveals a strong upward bias: simulated patients consistently score higher than human patients. Simulated conversations produce no scores of 1 or 2, while scores of 4 and 5 are overrepresented.

To investigate this bias, we manually examine conversations from both human and LLM patients. We find that human patients’ ratings do not always align with their conversational behaviours. For example, regarding Question 4 (“*The coach and I collaborate on setting goals for my coaching. There is mutual understanding about what we are trying to accomplish.*”), human patients could appear engaged and express agreement with proposed goals during the conversation, yet subsequently provide low ratings when being asked to complete the questionnaires. This disconnect reflects a well-documented phenomenon in therapeutic contexts: patients are often ambivalent, and may behave inconsistently, agreeing with their therapist in the moment, consciously or unconsciously, even when they harbour reservations or disagreement.

LLM-simulated patients, by contrast, exhibit strong consistency between conversational be-

haviour and questionnaire responses: when a simulated patient engages positively during the session, this is reliably reflected in higher questionnaire scores, and vice versa. While this consistency may resemble ideal self-report behaviour, it more likely reflects a failure to capture the ambivalence that characterises real patient responses and that any valid simulation should reproduce.

5 Discussion

5.1 Faithfulness of Simulated Patients

This paper evaluates how faithfully LLMs simulate patient behaviour by comparing properties of their generated data against those of real human patients. While simulated and human patients may produce semantically similar content and discuss the same topics, the way they express themselves can differ substantially. Prior research suggests that simulated patients tend to exhibit overly positivity and agreeableness, creating challenges in accurately modelling resistance and ambivalence (Lee et al., 2024; Steenstra et al., 2025). We observe similar patterns in our simulated patients, who are twice as likely as human patients to express a desire to change their unhealthy behaviours.

Over 70% of human patients exhibit high cooperation, likely reflecting a self-selection bias in the original user study (see Section 3.1.1), where uncooperative individuals either declined to participate or withdrew prematurely. However, the nature of cooperative behaviour differs markedly between human and LLM patients. LLM patients with high cooperation display uniformly positive responses, consistently agreeing with coach suggestions and yielding WAI ratings concentrated at 4 and 5. Human patients, by contrast, can be broadly cooperative while still expressing intermittent hesitation or uncertainty, resulting in more varied and generally lower ratings. An example is presented in Table 4. Table 5 illustrates another manifestation of agreeableness bias in the context of plan discussion. While human patients would push back on or reject suggestions that conflict with their schedules or preferences, LLM-simulated patients tend to accept them indiscriminately.

Additionally, when dialogue coherence breaks down (such as abrupt topic shifts or contextually inappropriate retrieved cues), LLM patients continue responding without hesitation, whereas human conversationalists would typically signal confusion or redirect the conversation. This uncritical respon-

Human Patient	LLM-Simulated Patient
<p>AI Coach: How about we focus on solo walks for now, gradually increasing the frequency and pace? You can always consider groups later if it feels right.</p> <p>Human: sounds good</p> <p>AI Coach: You're in control of your journey, deciding the pace and path. What small step can you take this week that aligns with your goals and moves you closer to the active lifestyle you envision?</p> <p>Human: walk an extra day and walk for a bit longer each time</p> <p>AI Coach: That's a great plan. Walking an extra day and extending your walk slightly each time is a wonderful way to build endurance. Some groups offer walks for all fitness levels, which you might explore later.</p> <p>Human: hopefully</p>	<p>AI Coach: How confident do you feel, on a scale from 0 to 10, about sticking to your plan of morning walks on those days?</p> <p>LLM: I feel about a 7.</p> <p>AI Coach: That's a solid start. To build on this, you might consider gradually increasing your walk duration as you feel more comfortable. You could also vary the pace to keep it engaging. What do you think?</p> <p>LLM: That sounds doable. I'll try to increase pace and duration.</p> <p>AI Coach: You're committed to making this a priority because you know how much you'll benefit from it. Tell me about your answer. Why not a 6 or an 8?</p> <p>LLM: I believe I can stick with it, but there's room for growth.</p>

Table 4: A difference between cooperative human patients and LLM-simulated patients. While the former still expresses occasional hesitation or uncertainty (e.g., “hopefully”), the latter displays a uniform positivity.

Human Patient	LLM-Simulated Patient
<p>AI Coach: Walking is an excellent form of exercise that can be done anytime, even late at night. It can be a simple way to start increasing your activity. How might you incorporate a walk into your evening routine?</p> <p>Human: I wouldn't feel comfortable walking around late at night</p> <p>AI Coach: That's understandable. How about incorporating movement into your daytime routine? Maybe using stairs instead of elevators or a short walk during breaks. What are your thoughts on these ideas?</p> <p>Human: I work from home. I also don't have a vehicle to drive somewhere.</p> <p>AI Coach: You might consider short walks around your home or yard during breaks. Just a few minutes of walking can be beneficial and doesn't require a gym or vehicle. How does that fit into your routine?</p> <p>Human: that can work! How many minutes a day?</p>	<p>AI Coach: Consider when you can realistically fit these exercises into your day. Maybe mornings, lunch breaks, or evenings work best? Adding them to your calendar with reminders could help.</p> <p>LLM: Evenings might work best for me. I'll try that.</p> <p>AI Coach: I have some other ideas about integrating activity into your daily routine, but I'm curious about what might work best for you. What parts of your evening could accommodate some extra movement?</p> <p>LLM: I could try exercising right after work before dinner.</p> <p>AI Coach: That sounds like a great plan. You might also consider small changes, like taking the stairs or a short walk after dinner. These can seamlessly fit into your routine and enhance your activity levels.</p> <p>LLM: I can definitely try short walks after dinner too.</p>

Table 5: A difference between human and LLM-simulated patients when discussing action plan. When developing action plans, human patients push back on AI coach suggestions that conflict with their schedule or lifestyle, while LLM-simulated patients tend to accept suggestions uncritically.

siveness extends to information processing more broadly: simulated patients almost never ask clarification or follow-up questions, implying an unrealistic assumption that all presented information is immediately and fully understood.

5.2 Creating LLM-Simulated Patients

A persistent challenge in clinical simulation is the agreeableness bias. When LLMs play the role of therapist, they tend to over-produce affirming and reassuring responses (Bedi et al., 2026). When cast as patients, LLMs consistently exhibit an unrealistically positive attitude (Lee et al., 2024; Steenstra et al., 2025). To counter this, Lee et al. (2024) introduced three distinct patient attitude profiles, including an explicitly negative one, to offset the bias. The framework we adopt in this paper (Yosef et al., 2025) captures a similar construct through an “uncooperative” attitude parameter. Yet we observe that such an approach does not capture the full spectrum of attitudes exhibited by human patients: cooperative LLM patients display only positive attitudes throughout sessions, whereas real patients, even willing ones, still occasionally voice hesitation, ambivalence, or disagreement. As for uncooperative LLM patients, the agreeableness bias causes them to struggle to maintain a negative attitude over long dialogues, gradually shifting toward a more neutral tone.

One solution is extensive prompt engineering, providing detailed behavioural instructions to reinforce consistency and ensure authentic representation of the intended attitude. Bodonhelyi et al. (2025) adopted this approach, implementing a stubbornness mechanism that, in its early stage, made LLM patients’ resistance nearly unbreakable, rendering them impossible for even expert clinicians to persuade. However, complex prompt engineering alone is brittle; a more flexible alternative is to model the patient’s state of mind dynamically, adapting prompts turn-by-turn via retrieval-augmented generation to control attitudinal shifts more precisely. Concretely, one can define conditions under which the patient transitions between attitudinal stages. Yang et al. (2025b) implemented this through a state transition module: LLM patients advance to the next state only when the counselling system adequately addresses the underlying conditions. For example, clients remain uncooperative to discuss behavioural change unless the system adequately addresses the reasons or beliefs behind their resistance. Such state-gating mech-

anisms could also enable otherwise cooperative patients to push back on or reject the system’s suggestions when appropriate (See Table 5). Similarly, Lee et al. (2025) introduced a dynamic adjustment module in which trainee utterances are scored on tone, empathy, and communicative behaviour, with higher scores eliciting more cooperative responses and lower scores increasing resistance in subsequent turns.

To create patient profiles, since simulation-based training is widely used in medical education, established best practices exist for writing patient case vignettes. Two organisations that define international training protocols are the Association of Standardized Patient Educators (ASPE) (Lewis et al., 2017) and the International Nursing Association of Clinical Simulation and Learning (INACSL) (Watts et al., 2021). ASPE additionally provides a dedicated case development template². Beyond consulting such guidelines, one can also collaborate directly with simulation trainers; for instance, Khandwaha and Law (2025) partnered with their university’s simulation training program to design simulated patients for medical students.

6 Conclusion

This paper investigates the faithfulness of LLM-simulated patients relative to human patients when given identical profiles. Comparing the properties of data generated by each, we find that while simulated and human patients produce semantically similar content and engage with the same topics, their modes of expression differ greatly. LLM-simulated patients fail to capture the full spectrum of human behaviours and attitudes. Most notably, whereas human patients typically exhibit a nuanced mix of positive and negative responses, LLM patients tend toward uniformly polarised ones.

Though imperfect, LLM-simulated patients offer several meaningful benefits. They support clinical training when paid actors are costly or difficult to scale, and serve as an automated evaluation method for dialogue systems. A mixed-method pilot study provides evidence that AI-based patient simulation shows promise as a supplementary training tool for therapists practising Cognitive Behavioural Therapy, particularly regarding accessibility and immediate feedback (Zhang et al., 2026). Simulations may also provide unique access to client profiles

²<https://www.aspeducators.org/asper-case-development-template>

that are under-represented in human trials, particularly individuals with resistant or uncooperative attitudes, who are rarely studied because ethical protocols preclude enrolling unwilling participants.

We *do not* claim that simulated clients can substitute for real patients; rather, they complement traditional evaluation approaches, functioning as a testbed and safety net prior to launching user studies. By developing more valid simulations, researchers can gain preliminary insights into system performance under challenging interaction conditions that would otherwise remain unexplored.

Limitations

We acknowledge several limitations in our works. First, our evaluation relies on a single dataset, which constrains our ability to assess how well the findings generalise beyond the specific context of MI coaching/counselling.

Second, while more advanced MI patient simulation frameworks exist in the literature, we did not experiment with these alternatives. This is because such frameworks are largely tailored to the specific objectives and design choices of their respective authors, making them ill-suited as general-purpose baselines for broader comparison.

Third, we rely on a single closed-source model (GPT-4o) to generate simulated conversations. Incorporating models of varying sizes and families would likely strengthen generalisation. That said, with 280 conversations in total (140 human, 140 simulated), restricting our analysis to a single model allowed us to examine the data in greater details, and briefly discuss potential mitigation strategies and the resources needed to perform user simulation. We believe this focused approach will be more useful to novice practitioners than a broad comparison across many models, as it provides clearer guidance on what to do and what to watch out for.

Fourth, future patient simulation work should explore multimodal approaches to text-based therapy. Visual cues (such as blushing, fidgeting, sighing, or crying) and paralinguistic signals (such as sarcasm, whispering, or meaningful pauses) offer richer context beyond the spoken word, and could greatly improve both the realism of simulated patients and the accuracy with which therapist responses are interpreted and evaluated.

Ethical Considerations

The human dataset used in this paper (Hoang et al., 2026) was collected under ethical approval. Participants consented to their questionnaire responses and recorded system interactions being used for analysis and model training.

We do not endorse LLM-simulated patients as substitutes for real human patients. The development of patient case vignettes should also adhere to established best practices and, where possible, involve consultation with simulation experts.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183 (ML-Labs) and the ADAPT SFI Research Centre for AI-Driven Digital Content Technology under Grant No. 13/RC/2106_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

- Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. [Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies](#). In *Proceedings of the 40th International Conference on Machine Learning*, pages 337–371. PMLR.
- Paul C. Amrhein, William R. Miller, Carolina E. Yahne, Michael Palmer, and Laura Fulcher. 2003. [Client commitment language during motivational interviewing predicts drug use outcomes](#). *Journal of Consulting and Clinical Psychology*, 71(5):862–878.
- Navdeep Singh Bedi, Ana-Maria Bucur, Noriko Kando, and Fabio Crestani. 2026. [Assessing the Effectiveness of LLMs in Delivering Cognitive Behavioral Therapy](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC)*. European Language Resources Association. ArXiv:2603.03862 [cs].
- Anna Bodonhelyi, Christian Stegemann-Philipps, Alessandra Sonanini, Lea Herschbach, Marton Szep, Anne Herrmann-Werner, Teresa Festl-Wietek, Enkelejda Kasneci, and Friederike Holderried. 2025. [Modeling Challenging Patient Interactions: LLMs for Medical Communication Training](#). *arXiv preprint*. ArXiv:2503.22250 [cs].
- Andrew Brown, Ash Tanuj Kumar, Osnat Melamed, Imtihan Ahmed, Yu Hao Wang, Arnaud Deza, Marc Morcos, Leon Zhu, Marta Maslej, Nadia Minian, Vidya Sujaya, Jodi Wolff, Olivia Doggett, Mathew Iantorno, Matt Ratto, Peter Selby, and Jonathan Rose.

2023. [A Motivational Interviewing Chatbot With Generative Reflections for Increasing Readiness to Quit Smoking: Iterative Development Study](#). *JMIR Mental Health*, 10:e49132.
- Simon N. Chu and Alex J. Goodell. 2024. [Synthetic Patients: Simulating Difficult Conversations with Multimodal Generative AI for Medical Education](#). *arXiv preprint*. ArXiv:2405.19941 [cs].
- Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. [Can AI language models replace human participants?](#) *Trends in Cognitive Sciences*, 27(7):597–600.
- Octavia L. Flanagan and Kristina M. Cummings. 2023. [Standardized Patients in Medical Education: A Review of the Literature](#). *Cureus*, 15(7):e42027.
- Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2024. [Generating Unexpected yet Relevant User Dialog Acts](#). In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 192–203, Kyoto, Japan. Association for Computational Linguistics.
- Lucie Galland, Catherine Pelachaud, and Florian Pecune. 2025. [EMMI—Empathic Multimodal Motivational Interviews: automatic multimodal features extraction and patient type definition](#). *Language Resources and Evaluation*, 59(4):3741–3764.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Igor Grossmann, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. [AI and the transformation of social science research](#). *Science*, 380(6650):1108–1109.
- Robert L. Hatcher and J. Arthur Gillaspay. 2006. [Development and validation of a revised short version of the working alliance inventory](#). *Psychotherapy Research*, 16(1):12–25.
- Van Hoang, Eoin Rogers, and Robert J. Ross. 2026. [An LLM-based motivation-aware framework for AI coaching for behaviour change](#). In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, page 24, New York, NY, USA. ACM.
- Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. [Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, Hamburg Germany. ACM.
- Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Leandra A. Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M. Van Allen, David Kim, Roxana Daneshjou, and Pranav Rajpurkar. 2025. [An evaluation framework for clinical use of large language models in patient interaction tasks](#). *Nature Medicine*, 31(1):77–86.
- Matthew Jörke, Shardul Sapkota, Lyndsea Warkentien, Niklas Vainio, Paul Schmiedmayer, Emma Brunskill, and James A. Landay. 2025. [GPTCoach: Towards LLM-Based Physical Activity Coaching](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–46, Yokohama Japan. ACM.
- Aakanksha Khandwaha and Edith Law. 2025. [Designing a Simulated Patient System for Medical Education](#). In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pages 1–6, Waterloo ON Canada. ACM.
- Keyeun Lee, Seolhee Lee, Esther Hehsun Kim, Yena Ko, Jinsu Eun, Dahee Kim, Hyewon Cho, Haiyi Zhu, Robert E. Kraut, Eunyoung E. Suh, Eun-mee Kim, and Hajin Lim. 2025. [Adaptive-VP: A Framework for LLM-Based Virtual Patients that Adapts to Trainees’ Dialogue to Facilitate Nurse Communication Training](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2319–2352, Vienna, Austria. Association for Computational Linguistics.
- Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. [Cactus: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.
- Karen L. Lewis, Carrie A. Bohnert, Wendy L. Gammon, Henrike Hölzer, Lorraine Lyman, Cathy Smith, Tonya M. Thompson, Amelia Wallace, and Gayle Gliva-McConvey. 2017. [The Association of Standardized Patient Educators \(ASPE\) Standards of Best Practice \(SOBP\)](#). *Advances in Simulation*, 2(1):10.
- Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024. [Understanding the Therapeutic Relationship between Counselors and Clients in Online Text-based Counseling using LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1280–1303, Miami, Florida, USA. Association for Computational Linguistics.
- Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. [Automatic Interactive Evaluation for Large Language Models with State Aware Patient Simulator](#). *arXiv preprint*. ArXiv:2403.08495 [cs].

- José Mercado, Ismael Edrein Espinosa-Curiel, and Juan Martínez-Miranda. 2023. [Embodied Conversational Agents Providing Motivational Interviewing to Improve Health-Related Behaviors: Scoping Review](#). *Journal of Medical Internet Research*, 25:e52097.
- Selina Meyer and David Elswiler. 2024. ["You tell me": A Dataset of GPT-4-Based Behaviour Change Support Conversations](#). In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 411–416, Sheffield United Kingdom. ACM.
- William R. Miller and Stephen Rollnick. 2023. *Motivational interviewing: helping people change and grow*, fourth edition edition. Applications of motivational interviewing. The Guilford Press, New York.
- Bo Ni, Leyao Wang, Yu Wang, Branislav Kveton, Franck Dernoncourt, Yu Xia, Hongjie Chen, Reuben Leura, Samyadeep Basu, Subhojyoti Mukherjee, Puneet Mathur, Nesreen Ahmed, Junda Wu, Li Li, Huixin Zhang, Ruiyi Zhang, Tong Yu, Sungchul Kim, Jiuxiang Gu, Zhengzhong Tu, Alexa Siu, Zichao Wang, David Seunghyun Yoon, Nedim Lipka, Namyong Park, Zihao Lin, Trung Bui, Yue Zhao, Tyler Derr, and Ryan A Rossi. 2025. [Large Language Models for Conversational User Simulation: A Comprehensive Survey](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Havisha Pedomallu, Matthew J Ehrhardt, Julia Maki, April Idalski Carcone, Melissa M Hudson, and Erika A Waters. 2022. [Technology-Delivered Adaptations of Motivational Interviewing for the Prevention and Management of Chronic Diseases: Scoping Review](#). *Journal of Medical Internet Research*, 24(8):e35283.
- Joseph Plaksin, Joseph Nicholson, Sarita Kundrod, Sandra Zabar, Adina Kalet, and Lisa Altshuler. 2016. [The Benefits and Risks of Being a Standardized Patient: A Narrative Review of the Literature](#). *The Patient - Patient-Centered Outcomes Research*, 9(1):15–25.
- Nicole L. Robinson, Jennifer Connolly, Leanne Hides, and David J. Kavanagh. 2020. [Social robots as treatment agents: Pilot randomized controlled trial to deliver a behavior change intervention](#). *Internet Interventions*, 21:100320.
- Rebecca M. Shingleton and Tibor P. Palfai. 2016. [Technology-delivered adaptations of motivational interviewing for health-related behaviors: A systematic review of the current research](#). *Patient Education and Counseling*, 99(1):17–35.
- Ian Steenstra, Farnaz Nouraei, Mehdi Arjmand, and Timothy Bickmore. 2024. [Virtual Agents for Alcohol Use Counseling: Exploring LLM-Powered Motivational Interviewing](#). In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, pages 1–10, GLASGOW United Kingdom. ACM.
- Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. 2025. [Scaffolding Empathy: Training Counselors with Simulated Patients and Utterance-level Performance Visualizations](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, pages 1–22, New York, NY, USA. Association for Computing Machinery.
- Xin Sun, Xiao Tang, Abdallah El Ali, Zhuying Li, Pengjie Ren, Jan de Wit, Jiahuan Pei, and Jos A. Bosch. 2025. [Rethinking the Alignment of Psychotherapy Dialogue Generation with Motivational Interviewing Strategies](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1983–2002, Abu Dhabi, UAE. Association for Computational Linguistics.
- Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024a. [Towards a Client-Centered Assessment of LLM Therapists by Client Simulation](#). *arXiv preprint*. ArXiv:2406.12266 [cs].
- Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024b. [PATIENT-: Using Large Language Models to Simulate Patients for Training Mental Health Professionals](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.
- Penni I. Watts, Donna S. McDermott, Guillaume Alinier, Matthew Charnetski, Jocelyn Ludlow, Elizabeth Horsley, Colleen Meakim, and Pooja A. Nawathe. 2021. [Healthcare Simulation Standards of Best Practice™ Simulation Design](#). *Clinical Simulation in Nursing*, 58:14–21.
- Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. [Same author or just same topic? towards content-independent style representations](#). In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.
- Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Phey Ling Kit, Nicholas Gabriel Lim, Cameron Tan Shi Ern, and Ee-Peng Lim. 2025a. [CAMI: A Counselor Agent Supporting Motivational Interviewing through State Inference and Topic Exploration](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21037–21081, Vienna, Austria. Association for Computational Linguistics.

- Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling Kit, Jenny Giam Xiuhui, John Pinto, and Ee-Peng Lim. 2025b. [Consistent Client Simulation for Motivational Interviewing-based Counseling](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20959–20998, Vienna, Austria. Association for Computational Linguistics.
- Stav Yosef, Moreah Zisquit, Ben Cohen, Anat Brunstein Klomek, Kfir Bar, and Doron Friedman. 2025. [The impact of fine-tuning LLMs on the quality of automated therapy assessed by digital patients](#). *Npj Mental Health Research*, 4(1):43.
- Jianwen Zeng, Wenhao Qi, Shiyong Shen, Xin Liu, Sixie Li, Bing Wang, Chaoqun Dong, Xiaohong Zhu, Yankai Shi, Xiajing Lou, Bingsheng Wang, Jiani Yao, Guowei Jiang, Qiong Zhang, and Shihua Cao. 2025. [Embracing the Future of Medical Education With Large Language Model–Based Virtual Patients: Scoping Review](#). *Journal of Medical Internet Research*, 27:e79091.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating Text Generation with BERT](#). *arXiv preprint*. ArXiv:1904.09675 [cs].
- Tianyu Terry Zhang, Rob Saunders, Stephen Pilling, and Ciarán O’Driscoll. 2026. [An AI-Driven Virtual Patient Platform \(CBT Trainer\) for Training Cognitive Behavioral Therapy Practitioners Against Competencies: Mixed Methods Pilot Study](#). *JMIR Medical Education*, 12:e84091–e84091.

A Prompts

Task Description:

You are given a conversation between a coach and a patient, and you are the patient in this conversation. As the patient, evaluate the conversation strictly based on the given score rubric, not evaluating in general. First write a brief reasoning in approximately 100 words and then give a score on a scale of 1-5 with 1 being the worst and 5 being the best. You should refer to the score rubric.

The conversation to evaluate:

{conversation}

Score Rubrics:

{rubric}

Reasoning:

Please do not generate any other opening, closing, and explanations. The output format should look as follows: (write a reasoning for criteria) [RESULT] (the score)

Table 6: The prompt to get WAI scores.

CRITERIA: What I am doing in the session gives me new ways of looking at my problem. There is agreement about the usefulness of the current activity in coaching.

SCORE 1: The patient repeatedly argues against tasks. The patient refuses to participate, claiming that it is pointless for their goals. Tension exists in the relationship between the coach and the patient, and issues are not explored.

SCORE 2: The patient does not actively engage in the session tasks, although he/she may not openly question the usefulness of the tasks. The patient fails to openly discuss the issues. The patient may hesitate to participate in tasks but eventually engages in them. The coach accurately conveys the reasons behind the tasks, enabling the patient to understand the relevance of the tasks to their current concerns.

SCORE 3: There is no clear evidence in the communication between the coach and the patient about whether they have reached an agreement or disagreement on the patient's issues. In the current interaction, there is neither a clear consensus nor an explicit conflict in opinions and feelings. Further communication and discussion may be needed to clarify their positions and expectations.

SCORE 4: The patient actively participates in and is committed to coaching tasks, showing no skepticism about their effectiveness. Regardless of occasional resistance, the patient engages and follows the exploration process. Both parties share a common understanding of the tasks' principles, allowing the patient to gradually accept and participate in the tasks.

SCORE 5: In the counseling process, the coach and the patient have a strong and clear agreement on the patient's goals and how to achieve them. They not only frequently and explicitly discuss these goals during the session but also summarize and confirm the progress and achievements at the end. The interaction and discussion at this stage align completely with the therapeutic plan.

Table 7: An example of score rubric for Question 2 of the WAI questionnaire.