SimPatient: Emotionally Realistic Simulated Patients for Counselor Empathy Training

Ian Steenstra*, Farnaz Nouraei*, Nishtha Sawhney*
*Northeastern University

***Abstract*—This research explores the design and application of emotionally realistic simulated patients for counselor empathy training. Counselor training often relies on didactic instruction and role-playing with standardized patients, which struggles to replicate the emotional complexity of real-world client interactions. This study introduces SimPatient, a system leveraging Large Language Models (LLMs) to create an embodied virtual patient using the Furhat robot, allowing novice counselors to practice their skills in a simulated environment. Two versions of SimPatient were tested: a control with basic dialogue responses and an intervention incorporating appraisal-based emotion modeling and nonverbal behaviors (NVB) generated via emotion-coping mechanisms. A small between-subjects study (n=6) compared the perceived emotional realism of the two versions and their impact on counselor empathy, measured using the Perth Empathy Scale pre- and post-interaction. While the emotion model enhanced perceived realism, especially regarding non-verbal behaviors (NVBs), it did not significantly increase empathy. In fact, empathy scores significantly decreased in the intervention group. This suggests that simply adding emotional responses may not be sufficient to improve empathy and highlights the complex interplay between realism, patient resistance, and counselor response. Further research with larger samples and refined implementations is needed to fully understand the potential of emotionally realistic simulated patients for empathy training.**

***Index Terms*—Large Language Models, Social Robots, Embodied Interaction, Appraisal Theory, Social Skills Training**

I. Introduction

Empathy is a crucial skill for counselors, enabling them to understand and respond effectively to clients' emotional experiences [1]. Developing and assessing this skill in novice counselors, particularly when faced with resistant clients, presents a significant challenge. Traditional training methods, such as didactic instruction and role-playing with standardized patients, often struggle to replicate the emotional complexity, dynamic nature, and resistance commonly encountered in real-world client interactions, especially in challenging contexts like substance misuse counseling [2]. Cognitive models of addiction highlight the interplay of cognitive biases and emotional responses to alcohol cues, which can trigger cravings, impulsive behaviors, and resistance to treatment [3], [4]. This limitation motivates the exploration of innovative training tools that provide more engaging, realistic, and challenging learning experiences. Virtual patient simulations offer a promising approach, providing a safe and repeatable environment for practicing these complex skills. However, many existing systems lack the emotional realism required to elicit genuine empathic responses, especially when simulating resistant patients [5], [6]. Recent advances in Large Language Models (LLMs) and embodied conversational agents offer the potential to bridge this gap.

This research investigates SimPatient, a novel system for counselor empathy training that leverages LLMs to create an emotionally realistic simulated patient embodied by the Furhat robot. Our initial test domain focuses on simulating highly resistant alcohol misuse patients, recognizing that counselor empathy is frequently tested when interacting with this challenging patient population. Unlike previous systems with narrowly scripted scenarios, SimPatient utilizes LLMs to generate dynamic and contextually appropriate responses, informed by alcohol-related beliefs, desires, and intentions, aiming to effectively simulate the irrational and biased emotional processing often observed in individuals struggling with alcohol use. This study focuses specifically on incorporating appraisal-based emotion modeling and nonverbal behaviors (NVB) into the simulated patient to enhance realism and elicit more authentic empathic reactions from trainees. By integrating a computational model of appraisal theory [7] with the expressive capabilities of the Furhat robot, SimPatient aims to create a more immersive and engaging training environment. This work addresses the challenges faced by counselor training programs in providing consistent, cost-effective access to standardized patients capable of accurately portraying specific symptoms, patient resistance, and the effects of substance use disorders [2].

Two versions of SimPatient were developed and tested: a control condition with basic dialogue and an intervention condition incorporating the emotion model and NVBs. A between-subjects study (n=6) compared the perceived emotional realism of both versions and their impact on counselor empathy, as measured by the Perth Empathy Scale [8]. This study addresses the following research questions:

- **R1:** What is the impact of appraisal-based emotional reasoning on the perceptions of novice counselors about simulated patient emotional realism?
- **R2:** To what extent does overall empathy, as measured by the Perth Empathy Scale, change after interaction with an emotionally-enabled simulated patient?

II. Background – Appraisal Theory

Appraisal theory is a cognitive framework that explains how emotions arise from an individual's evaluation of a situation. This evaluation is based on several dimensions that shape the type and intensity of emotional responses. These dimensions

include *relevance* (whether the situation affects personal goals or well-being), *desirability* (whether the outcome is perceived as good or bad), and *expectedness* (how anticipated the outcome is). Additionally, *likelihood* assesses the probability of an event occurring, while *controllability* evaluates whether the outcome can be influenced or managed. *Changeability* reflects how adaptable or modifiable the situation is, and *causal attribution* identifies who or what is responsible for the outcome [9].

According to Lazarus's seminal work on appraisal theory, these evaluations occur in two stages: *primary appraisal* and *secondary appraisal*. In the primary appraisal, individuals assess the *relevance* and *desirability* of a situation, determining whether it poses a threat, offers a benefit, or is irrelevant. In the secondary appraisal, they evaluate *controllability*, *changeability*, and *causal attribution* to determine their capacity to cope with the situation. These cognitive evaluations dynamically shape emotional experiences, demonstrating that emotions are adaptive responses aligned with personal goals and coping abilities [9].

III. Related Work

A. Appraisal Theory & LLMs

Appraisal theory has increasingly become a valuable framework for modeling emotions in LLMs [10], [11]. Recent research has sought to explore how LLMs process emotions through this theoretical lens and how well their responses align with human emotional appraisals.

For example, Yongsatianchot et al. investigated LLMs' perception of emotions using appraisal theory by examining correlations between appraisals generated by LLMs and those generated by humans [10]. Their findings indicate that while LLMs show trends similar to humans in appraisal dynamics, they struggle to differentiate between specific types of scenarios, such as aversive versus loss scenarios. Furthermore, LLMs exhibit notable deviations from human responses in key appraisal dimensions like controllability and coping.

Interestingly, when LLMs like ChatGPT and GPT-4 are instructed to simulate a particular emotional state, such as depression, their outputs align more closely with theoretical predictions [10]. This suggests that while LLMs are sensitive to the structure and phrasing of prompts, careful prompting strategies can elicit more accurate emotional appraisals.

Building on these insights, Croissant et al. introduced an appraisal-based chain-of-emotion architecture designed for affective language model agents used in gaming contexts [11]. Their approach incorporates appraisal prompting to determine an agent's emotional state before generating responses. Specifically, a two-step process is employed: the model first appraises the situation to generate an emotional description, which is then integrated into the prompt for subsequent response generation. This method improves the believability and emotional intelligence of the agents.

B. Standardized Patients

Traditionally, standardized patients–actors who allow providers to practice clinical skills by enacting real-life scenarios as a patient–have been widely used in healthcare to enhance medical education, primarily for communication and clinical skills. Prior research indicates that the use of standardized patients allows trainees to practice real-life scenarios in a controlled environment, fostering a sense of realism and emotional engagement that traditional educational methods may lack. Previous work has emphasized the effectiveness of standardized patients in medical training by showing that engaging with standardized patients leads to better acquisition of communication skills compared to didactic methods [12]. However, using standardized patients can be costly and time consuming, due to cumbersome recruiting, training, and standardization of their interactions [13]–[15].

C. Patient Simulation Systems

Patient simulation systems have undergone a transformation with the advent of technologies like chatbots and embodied conversational agents. Simulated patients offer a flexible training approach, allowing learners to interact with computerized patients capable of natural language processing and realistic emotional responses [16]. These systems offer advantages in cost-effectiveness and availability compared to using human standardized patients, while also enabling personalized and context-rich training scenarios.

Virtual reality (VR) has further broadened the possibilities of virtual patient simulations for medical communication training [17]. Lok et al. demonstrated the effectiveness of VR-based systems with virtual patients, providing immersive, life-sized clinical interactions for practicing skills such as delivering difficult news and conducting patient interviews. Medical students perceived these VR simulations as both authentic and valuable learning tools. The capacity for reflective practice and feedback within a low-risk environment underscores the scalability and cost-effectiveness of VR-based training as a viable alternative to traditional standardized patient methods. High-fidelity patient-focused simulations, integrating real human patients with simulators, have also been employed to develop communication and decision-making skills alongside technical proficiency [18]. Studies show that increased realism in simulations correlates with improved skill and knowledge retention, enhanced engagement, and greater clinical confidence [19].

Furthermore, research by Spear et al. [20] utilized simulated patients to train doctors in delivering difficult news. Their study, encompassing both single- and multiple-learner scenarios, demonstrated significant improvements in communication skills across both groups. Notably, their work highlighted the potential of interdisciplinary training by combining pediatricians and nurses within the same simulation.

The integration of human-like interaction in patient simulation systems has proven crucial for bridging the gap between theoretical knowledge and practical application in healthcare [18], [19], [21].

D. LLMs for Simulated Patient Dialogue

Recent advancements in LLMs have made them an attractive candidate to drive simulated patient dialogue and system reasoning for training purposes. To this end, CureFun utilizes multiple LLMs to simulate patients for medical diagnosis education, and offers automated feedback on trainee performance by analyzing medical scenarios to inform its responses [22]. Similarly, [23] introduce State-Aware Patient Simulator to bridge the gap between static medical knowledge assessments and dynamic clinical interactions [23]. Their approach allows for more realistic LLM evaluation in multi-turn doctor-patient simulations.

Despite impressive conversational abilities of LLMs, ensuring accurate and nuanced portrayals of specific patient populations requires expert input. [24] address this challenge with Roleplay-doh, a pipeline that enables domain experts to guide LLM simulations through elicited principles. This highlights the crucial role of expert feedback, especially in sensitive domains like mental health, by providing a mechanism for its direct incorporation into LLM prompting [24]. [25] explore the use of ChatGPT for the simulation of psychiatrists and patients, developing a dialogue system informed by psychiatrist input [25]. Their evaluation with real clinicians highlights the feasibility of LLM-powered chatbots in psychiatric scenarios while emphasizing the importance of careful prompt design for realistic and ethical interactions. Recent research demonstrates that inaccuracies in virtual patient responses can hinder clinicians' communication skills and empathy development during training simulations [26]. These discrepancies can disrupt engagement, evoke frustration, and reduce the overall effectiveness of the training. Ensuring virtual patient interactions are contextually appropriate and realistic is crucial for preserving the authenticity and educational value of LLM-driven systems.

The use of LLMs for social skill training, including conflict resolution and communication, presents a promising application [27], [28]. For example, [28] discuss a framework leveraging LLMs to create accessible and engaging social skills training environments [28]. Their AI Partner, AI Mentor framework combines experiential learning with realistic practice and tailored feedback, which aligns with SimPatient's goal of providing a safe space for practicing MI techniques. However, LLMs, especially those trained with reinforcement learning from human feedback as analyzed by Perez et al. [29], exhibit a positivity bias, tending towards agreeableness and a desire to comply with user requests. This raises challenges for simulating resistant or deceptive patients, scenarios crucial for robust training [30]–[32]. Further research is needed to overcome these limitations and accurately model the full spectrum of patient behaviors. Closest to our work is that of Wang et al., who proposed a simulated patient framework to enhance Cognitive Behavioral Therapy training [5]. In their work, patient cognitive models, such as emotional states and maladaptive cognitions grounded in the principles of Cognitive Behavioral Therapy, are integrated with LLMs to ensure high-fidelity simulated patient interactions.

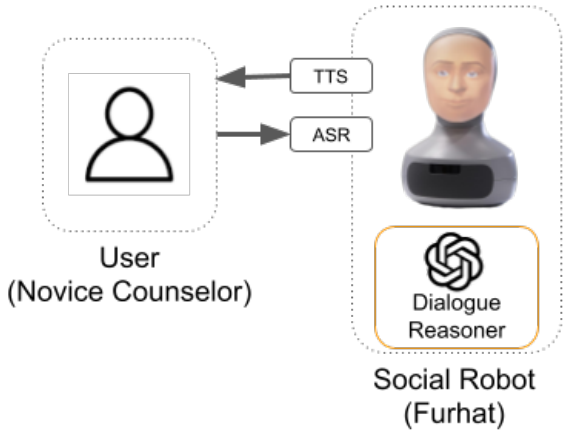

Fig. 1. Interaction Design

IV. System Design

Our SimPatient system includes two main components: An utterance generation module powered by GPT-4o drives the content of the interaction, while the Furhat social robot[1] is used to represent the patient by realizing those interactions through verbal and NVBs (Figure 1). We used Furhat's built-in speech synthesizer to convert LLM-generated utterances into voice. We created two versions of our system to investigate the impact of emotion on patient realism and counselor empathy. In the intervention version (emotion module enabled), the utterance generation module includes emotion reasoning that guides the generation of utterances, which Furhat utters during interactions. In this version, NVBs are also selected by the LLM and are realized using Furhat's gesture library, which includes template realizations of basic emotions such as happiness, surprise, and sadness, via smiling or nodding NVBs. In addition, a slight gaze aversion function renders the robot interaction more natural when talking to the user.

In the control version (emotion module disabled), the LLM does not perform emotional reasoning, and no facial expressions are displayed by Furhat. The robot operates in idle mode, tracking the user's face with head movements but without expressing emotions.

In both versions of the system, we provide the LLM with a contextual guide (task description) and patient attributes and persona (age, gender, personality traits, a description of alcohol use habits and attitudes toward change), which can be customized by the novice counselor users. We include our code (including prompts) for reproduciblity[2]

A. Emotion Module

An emotion module was created for use in the first version of our system. We based our emotion modeling approach on previous work on appraisal-based computational models of

[1]https://www.furhatrobotics.com
[2]https://github.com/IanSteenstra/FurhatPatient

emotions, due to its relevance to an alcohol problem scenario, and the explicit variables that can be defined in an LLM prompt (as opposed to numerical dimensions only). To this end, we adapted the computational model proposed in EMA [7], which includes important components such as attentional focus, appraisal, coping, and dynamic beliefs, desires and intentions. We also adapted the work by Yongsatianchot et al. [10] for minor prompt engineering tweaks.

SimPatient reasoner processes information on a turn by turn basis. Regardless of the turn, the LLM system prompt begins with the contextual guide and patient attributes described above. Initially, we also provide a seed for up to three beliefs, desires and intentions that the patient holds prior to the conversation. Keeping in mind this information the LLM is then asked to take a chain-of-thought approach [33] to derive appropriate patient emotions. In the rest of this section, we discuss how we used each concept to craft our LLM prompt. Prior to each step, the LLM was asked to first generate its reasoning about that step, which we found helpful in both guiding the model's (numerical) generations and debugging potential issues in the prompt.

a) Attention: Appraisal theories suggest that emotions are experienced through an assessment of events based on their subjective significance to the person. As a first step to this, our model should detect an event or situation that is important enough to the patient to form an appraisal frame for at the moment. From a technical perspective, we found that including this step in the chain of thought can reduce the inconsistency of the events being appraised within one turn, such that without this element, the LLM may generate the value of one appraisal variable with regards to, say, the patient's current health status, while another variable may be generated about the patient's relationship with the counselor.

b) Appraisal: An appraisal frame in our work includes 7 variables from EMA, namely relevance, desirability, likelihood, expectedness, causal attribution, controllability and changeability. To capture the intensity (importance) of each appraisal, we used a Likert scale for the LLM to choose from, where 0 represents very small and 5 represents a very large effect. In our prompt, we emphasized that these appraisals are subjective to the patient in question and do not have to make sense to others. Our experiments qualitatively showed that such emphasis allows the language model to break out of its fidelity to factual and rational reasoning, and represent extremes such as alcohol dependency and negative emotions.

c) Pleasure-Arousal-Dominance(PAD): While appraisals capture the significance of the event, we hypothesized that a numerical representation of pleasure, arousal, and dominance can narrow the space of emotions for the model, increasing its capability to assign accurate emotion labels in the next step. We therefore asked them model to score the patient's emotion on these using a value between -1 and 1 for each dimension.

d) Emotion Labels: Based on the predictions so far, the models picks up to three emotion lables that the patient is likely to experience at this turn. We provided the LLM with the 26-emotion list in the Emotic dataset [34] to capture the emotional subtleties of alcohol use situations. The labels in this dataset are a categorization of more fine-grained emotions, thus reducing the vast space of emotion to an extent managable by an LLM predictor. Examples of emotion labels often predicted by our system in the SimPatient experiments include: disquietment, engagement, fear, annoyance and disapproval.

e) Coping Potential: A distinctive part of our emotion modeling approach compared to other AI-based approaches [10], [11], [35] is modeling coping dynamics, and how they influence utterance generation and the conversation trajectory. Coping relies on appraisal to identify significant features of the person–environment relationship and to assess the potential to maintain or overturn these features, also called coping potential [7]. In our model, the LLM scores the coping potential of the appraisal frame related to the event on a scale of 0 to 5.

f) Coping Approach: Based on the coping potential and other previously generated values, the model then chooses a coping approach from among a list of 16 problem- and emotion-focused approaches, which we collected from [7] and [36].

g) Beliefs, Desires and Intentions(BDI): Once the coping approach is generated, the BDI list pertaining to the patient is updated using the new appraisal and coping information, closing the loop of the emotion model. We found that a contextually-relevant BDI seed (which we manually provided to the LLM at the beginning of the interaction), combined with the dynamic BDI updates after each turn, leads to a significantly more realistic (and potentially resistant to change) patient and, to some extent, mitigates the issue of LLM submissiveness, i.e., LLMs' tendency to produce agreeable and instruction-following output rather than realistically simulate a human [28], [29].

V. Experiment

This study used a between-subjects design to evaluate SimPatient's emotion module's impact on perceived realism and counselor empathy. Participants interacted with a simulated patient, embodied by the Furhat robot platform, and were assigned via block randomization to either an intervention group (emotion module enabled) or a control group (emotion module disabled). All participants interacted with the same resistant patient persona, "Alex," characterized by alcohol misuse, reluctance to change, and irritability due to being forced into counseling by his friends. Alex's persona was defined by characteristics including: enjoying partying but experiencing frequent negative consequences; being resistant to addressing his alcohol use; being a male, Asian, PhD student with an INFJ-A personality type; and having scores (on scales of 1-10) of 5 for Control Level (representing perceived ability to regulate thoughts, emotions, and actions related to alcohol), 1 for Self-Efficacy (reflecting confidence in resisting cravings and achieving recovery goals), 1 for Awareness (representing insight into thoughts, feelings, and behaviors related to alcohol use), and 6 for Reward (reflecting the degree to which alcohol

triggers cravings and automatic behaviors). These characteristics informed the LLM's responses, aiming to create a consistent and challenging interaction. The procedure took 20-minutes, and involved participants completing a pre-interaction survey, engaging in a 10-minute interaction with SimPatient (Alex) in their assigned condition, and then completing a post-interaction survey and a semi-structured interview. All participants provided informed consent after receiving an explanation of the study's purpose and procedures.

A. Measures

Two primary measures were used in this study: the Perth Empathy Scale [8] and a set of questions designed to assess the perceived emotional realism of the simulated patient. The Perth Empathy Scale was administered both before and after the interaction with SimPatient to assess any changes in participant empathy levels. The perceived emotional realism questions, along with a semi-structured interview exploring participant experiences, were administered only after the interaction.

1) Perth Empathy Scale: The Perth Empathy Scale [8] is a 20-item self-report measure of empathy, assessing both cognitive and affective components across positive and negative valences [37]. The Cognitive Empathy (CE) subscale measures the ability to recognize others' emotions, while the Affective Empathy (AE) subscale measures the capacity to vicariously experience those emotions. Each item is rated on a Likert scale from 1 (Almost Never) to 5 (Almost Always). Higher total scores indicate greater empathy. A total composite score represents overall empathy ability; separate subscale scores can be calculated for negative/positive cognitive empathy and negative/positive affective empathy. This study used the total empathy composite score, as well as the CE and AE subscales.

2) Perceived Emotional Realism: Three questions assessed participants' perceptions of the simulated patient's emotional realism after interacting with SimPatient:

1) **Emotion Realism:** *"Considering the patient history and their dialogue during the interaction, how realistically did the virtual patient display emotion?"* (1-5 Likert scale from Not realistic at all to Extremely realistic)
2) **Non-Verbal Realism:** *"Given the context and scenario, how realistic were the non-verbal cues of the virtual patient?"* (1-5 Likert scale from Not realistic at all to Extremely realistic)
3) **Complex Emotion:** *"How complex were the emotions displayed by the virtual patient?"* (1-5 Likert scale from Not complex at all to Highly complex)

These questions aimed to capture different facets of emotional realism, including the appropriateness of emotional responses, the believability of NVBs, and the depth and nuance of emotions expressed by the simulated patient.

B. Results

Participants: Six participants were recruited for this study. Demographics were as follows: 83.33% aged 25-34 and 16.67% aged 35-44; 50% male and 50% female; occupations included student (66.67%), engineer (16.67%), and developer (16.67%); racial identification was White (50%), Asian (33.33%), or prefer not to say (16.67%); 83.33% identified as Not Hispanic or Latino/a/e and 16.67% preferred not to say; AI interaction frequency was daily (50%), weekly (33.33%), or rarely (16.67%). Prior experience with AI systems included: chatbots (83.33%), virtual assistants (100%), VR experiences (33.33%), gaming AI (33.33%), and virtual simulation bots (33.33%). Lastly, participants indicated high openness to using AI for learning when asked *"How open are you to engaging with virtual agents or AI systems for learning new skills?"*, on a 1-5 Likert scale from Not open at all to Very open (M = 4.5 out of 5, SD = 1.12).

Perth Empathy Scale Table I presents the mean and standard deviation for the total empathy score, CE, and AE for both groups, pre- and post-interaction. A Shapiro–Wilk test was used to confirm normality. Paired t-tests were used to assess within-group changes from pre- to post-interaction. For the control group, no significant changes were observed (Total Empathy: $t(2) = 1.15$, $p = 0.37$; CE: $t(2) = 1.51$, $p = 0.27$; AE: $t(2) = 0.65$, $p = 0.58$). One the contrary, the intervention group showed a significant decrease in total empathy (Total Empathy: $t(2) = -7.00$, $p = 0.02^{*}$; CE: $t(2) = -1.73$, $p = 0.23$; AE: $t(2) = -4.00$, $p = 0.06$). Using unpaired t-tests, we found no significant between-group differences in (post-pre) change scores (Total Empathy: $t(4) = 1.88$, $p = 0.13$; CE: $t(4) = 1.98$, $p = 0.12$; AE: $t(4) = 1.5$, $p = 0.21$).

TABLE I
PERTH EMPATHY SCALE SCORES

	Control		Intervention	
	Pre (M, SD)	Post (M, SD)	Pre (M, SD)	Post (M, SD)
Total Empathy	67.33 (6.85)	71.00 (11.31)	78.67 (8.81)	76.33 (8.34)
CE	35.67 (3.30)	38.33 (5.44)	41.33 (2.87)	40.33 (2.05)
AE	31.67 (3.86)	32.67 (5.91)	37.33 (6.13)	36.00 (6.38)

Perceived Emotional Realism Table II presents the mean and standard deviation for each realism question. A Shapiro–Wilk test was used to confirm normality. Unpaired t-tests comparing the control and intervention groups revealed no significant differences (Emotion Realism: $t(4) = 1$, $p = 0.37$; Non-Verbal Realism: $t(4) = 0.5$, $p = 0.64$; Complex Emotion: $t(4) = 1$, $p = 0.37$).

TABLE II
PERCEIVED EMOTIONAL REALISM SCORES

	Control (M, SD)	Intervention (M, SD)
Emotion Realism	**4.33 (0.47)**	4.00 (0.00)
Non-Verbal Realism	3.00 (0.82)	**3.33 (0.47)**
Complex Emotion	3.00 (0.82)	**3.67 (0.47)**

Observations & Semi-Structured Interviews The first author, who facilitated the study and conducted the exit semi-structured interviews, performed a rapid thematic analysis of their observation and interview notes. Several themes emerged from the rapid thematic analysis of observation and interview data.

A notable difference appeared in patient response timing between the control and intervention groups. The simulated patient in the control condition responded in one second or less, while responses in the intervention condition took 2-5 seconds. This difference was attributed to the additional processing required for the emotion module in the intervention condition. Participants in the intervention group also noted a greater frequency of NVBs from the simulated patient. While sometimes perceived as excessive, these NVBs were generally considered realistic and appropriate to the context. Across both groups, participants generally found the simulated patient to be realistic and resistant to the counseling, with one participant even reporting a sense of connection with the virtual patient. Finally, some technical limitations were identified, primarily related to the automated speech recognition (ASR) system, which struggled with accents, delays, and speech cutoffs. Challenges with facial tracking, due to lighting, framing, and distance, were also noted. The ASR system also required participants to maintain speaking continuity without pauses for optimal function.

VI. Discussion

A. RQ1: Impact of Emotional Reasoning on Perceived Realism

Our first research question explored the impact of appraisal-based emotional reasoning on perceived realism. While the quantitative results did not reveal statistically significant differences between the control and intervention groups on any of the realism subscales, the qualitative data suggests a nuanced impact. Participants in the intervention group noticed and commented on the increased presence of NVBs, which, despite sometimes being perceived as excessive, were generally deemed contextually appropriate. This aligns with prior work in affective computing, which suggests that incorporating appraisal theories into computational models can lead to more believable and emotionally intelligent agents [10], [11], [38]. This indicates that the appraisal-based emotion model may have influenced the perceived realism, primarily through the generation of NVBs. The lack of significant differences in the quantitative measures could be attributed to the small sample size and the novelty of the interaction paradigm, potentially impacting participants' ability to discern subtle differences in emotional expression.

B. RQ2: Empathy Change After Interaction with Emotionally-Enabled Patient

Our second research question investigated whether interacting with an emotionally-enabled simulated patient would lead to changes in overall empathy. Contrary to our hypothesis, no significant increase in empathy was observed in the intervention group. In fact, the intervention group showed a significant *decrease* in total empathy. This unexpected result warrants further investigation and could be due to several factors. First, the simulated patient's resistance, driven by the underlying persona and potentially amplified by the emotion model's influence on dialogue, may have evoked negative emotional responses in participants, potentially decreasing their empathic response. This highlights the complex interplay between patient behavior and counselor empathy, emphasizing that realistic emotional responses alone may not be sufficient to enhance empathy. It is possible that specific training or guidance on how to manage challenging patient interactions is necessary to realize the full potential of emotionally realistic simulated patients for empathy training. Second, the short interaction time (10 minutes) may not have been sufficient to induce measurable changes in empathy, a trait often developed over longer periods and through repeated experiences [39]. Furthermore, the small sample size limits the generalizability of these findings.

C. Limitations and Future Work

This study has several limitations, primarily stemming from the small sample size (n=6), which significantly restricts the statistical power and generalizability of the findings. Future work should involve a larger and more diverse participant pool, including actual novice counselors, to assess SimPatient's effectiveness in a more ecologically valid setting. The relatively short interaction time (10 minutes) may have limited the potential impact on empathy development. Longer and more varied interaction scenarios could provide more opportunities for observing changes in empathy. Additionally, while this preliminary study utilized a convenience sample, future studies should prioritize a more representative sample, including controlling for factors like prior experience with alcohol and beliefs about addiction.

From a technical viewpoint, our proposed emotion module is complex and involves several steps of LLM reasoning, increasing interaction latency. Even though the present study did not show significant negative effects of latency, future work should include a more systematic approach to testing each step of our emotion model and pruning steps that may not be contributing to variations in output, thereby reducing cost associated with using the system.

The current study design can be improved by incorporating a larger sample size with diverse personas exhibiting varying levels of resistance, allowing for generalized comparisons and more comprehensive testing of the intervention's effectiveness. The technical limitations encountered with the ASR and facial tracking systems should also be addressed in future work to enhance the overall user experience and ensure reliable data collection.

Further research should explore the specific mechanisms through which appraisal-based emotion modeling influences counselor responses. Investigating how the model impacts communication patterns, emotional recognition, and intervention strategies could provide valuable insights for refining the system and maximizing its training potential. Exploring

the integration of other affective computing techniques, such as sentiment analysis and emotion recognition from speech, could further enhance the simulated patient's responsiveness and create even more nuanced interactions.

Finally, expanding the evaluation metrics beyond generalized empathy measures would provide a richer understanding of SimPatient's impact. Incorporating validated questionnaires specific to the training domain (e.g., alcohol and other drug abuse treatment) could offer a more targeted assessment of counselor skill development. Physiological measurements of empathy, such as electrodermal activity, could also be valuable additions, offering objective indicators of emotional responses during the interaction.

A planned second phase of this research will address some of these limitations. Expert counselors will analyze dialogue transcripts from the interactions, evaluating patient appraisals, the need for empathy in each turn, the counselor's empathic responses, and ideal responses. This expert analysis will provide a valuable benchmark for evaluating the system's performance and identifying areas for improvement. Furthermore, the experts will rate the accuracy of an automated empathy evaluation tool on a 1-10 scale for each dialogue turn, contributing to the development of objective metrics for assessing counselor empathy in simulated interactions. This two-phase approach will contribute valuable insights for improving future iterations of SimPatient and maximizing its effectiveness as a training tool for counselors.

VII. Conclusion

This study investigated the potential of using emotionally realistic simulated patients, powered by appraisal-based emotion modeling and embodied by the Furhat robot, for counselor empathy training. While the implemented emotion model influenced perceived realism, particularly in NVBs, it did not lead to a statistically significant increase in counselor empathy, and even resulted in a significant decrease in empathy in the intervention condition. This highlights the complex relationship between perceived realism, patient resistance, and counselor empathy, suggesting that simply adding emotional responses may not be sufficient to enhance empathy training. Future research with larger samples, refined technical implementations, and extended interaction paradigms is needed to fully explore the potential of emotionally realistic simulated patients and to identify the most effective strategies for integrating them into counselor training programs. Despite the limitations, this work contributes to the growing field of affective computing in healthcare and provides valuable insights for the development of more sophisticated and effective virtual patient training tools.

Acknowledgment

Credit Assessment:

- Ian: Coding (85%), Conducting Study (100%), Data Analysis (100%)
- Farnaz: Coding (15%), Prompt Engineering (100%), Background Reading: Appraisal Theory (50%), Related Work (50%)
- Nishtha: Metrics/Measures Selection (100%), Survey Creation (100%), Background Reading: Appraisal Theory (50%), Related Work (50%)

References

[1] K. M. Decker, *A study of relationships between counselor education, social justice advocacy competence, and likelihood to advocate*. Capella University, 2013.

[2] M. B. Madson, A. C. Loignon, and C. Lane, "Training in motivational interviewing: A systematic review," *Journal of substance abuse treatment*, vol. 36, no. 1, pp. 101–109, 2009.

[3] M. L. Copersino, "Cognitive mechanisms and therapeutic targets of addiction," *Current opinion in behavioral sciences*, vol. 13, pp. 91–98, 2017.

[4] E. L. Garland, B. Froeliger, and M. O. Howard, "Mindfulness training targets neurocognitive mechanisms of addiction at the attention-appraisal-emotion interface," *Frontiers in psychiatry*, vol. 4, p. 173, 2014.

[5] R. Wang, S. Milani, J. C. Chiu, S. M. Eack, T. Labrum, S. M. Murphy, N. Jones, K. Hardy, H. Shen, F. Fang, *et al.*, "Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals," *arXiv preprint arXiv:2405.19660*, 2024.

[6] S. Yosef, M. Zisquit, B. Cohen, A. K. Brunstein, K. Bar, and D. Friedman, "Assessing motivational interviewing sessions with ai-generated patient simulations," in *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pp. 1–11, 2024.

[7] S. C. Marsella and J. Gratch, "Ema: A process model of appraisal dynamics," *Cognitive Systems Research*, vol. 10, no. 1, pp. 70–90, 2009.

[8] J. D. Brett, R. Becerra, M. T. Maybery, and D. A. Preece, "The psychometric assessment of empathy: Development and validation of the perth empathy scale," *Assessment*, vol. 30, no. 4, pp. 1140–1156, 2023.

[9] R. S. Lazarus, *Emotion and Adaptation*. Oxford University Press, 1991.

[10] N. Yongsatianchot, P. G. Torshizi, and S. Marsella, "Investigating large language models' perception of emotion using appraisal theory," in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 1–8, IEEE, 2023.

[11] M. Croissant, M. Frister, G. Schofield, and C. McCall, "An appraisal-based chain-of-emotion architecture for affective language model game agents," *Plos one*, vol. 19, no. 5, p. e0301033, 2024.

[12] C. Lane and S. Rollnick, "The use of simulated patients and role-play in communication skills training: a review of the literature to august 2005," *Patient education and counseling*, vol. 67, no. 1-2, pp. 13–20, 2007.

[13] O. L. Flanagan and K. M. Cummings, "Standardized patients in medical education: a review of the literature," *Cureus*, vol. 15, no. 7, 2023.

[14] A. Whitaker, M. Quinn, S. Martins, A. Tomlinson, E. Woodhams, and M. Gilliam, "Motivational interviewing to improve postabortion contraceptive uptake by young women: development and feasibility of a counseling intervention," *Contraception*, vol. 92, no. 4, pp. 323–329, 2015.

[15] D. B. Swanson and C. P. van der Vleuten, "Assessment of clinical skills with standardized patients: state of the art revisited," *Teaching and learning in medicine*, vol. 25, no. sup1, pp. S17–S25, 2013.

[16] R. C. Hubal, P. N. Kizakevich, C. I. Guinn, K. D. Merino, and S. L. West, "The virtual standardized patient-simulated patient-practitioner dialog for patient interview training," in *Medicine Meets Virtual Reality 2000*, pp. 133–138, IOS Press, 2000.

[17] B. Lok, R. E. Ferdig, A. Raij, K. Johnsen, R. Dickerson, J. Coutts, A. Stevens, and D. S. Lind, "Applying virtual reality in medical communication education: current findings and potential teaching and learning benefits of immersive virtual patients," *Virtual Reality*, vol. 10, pp. 185–195, 2006.

[18] R. S. Bartlett, S. Bruecker, and B. Eccleston, "High-fidelity simulation improves long-term knowledge of clinical swallow evaluation," *American Journal of Speech-Language Pathology*, vol. 30, no. 2, pp. 673–686, 2021.

[19] R. Kneebone, D. Nestel, C. Wetzel, S. Black, R. Jacklin, R. Aggarwal, F. Yadollahi, J. Wolfe, C. Vincent, and A. Darzi, "The human face of simulation: patient-focused simulation training," *Academic Medicine*, vol. 81, no. 10, pp. 919–924, 2006.

[20] M. L. Spear, L. J. Rockstraw, and B. Bernstein, "Comparison of single versus multiple learners in standardized patient communication skills training in palliative care," *Journal of Interprofessional Education & Practice*, 2019.

[21] S. Sarker and B. Patel, "Simulation and surgical training," *International journal of clinical practice*, vol. 61, no. 12, pp. 2120–2125, 2007.

[22] Y. Li, C. Zeng, J. Zhong, R. Zhang, M. Zhang, and L. Zou, "Leveraging large language model as simulated patients for clinical education," *arXiv preprint arXiv:2404.13066*, 2024.

[23] Y. Liao, Y. Meng, Y. Wang, H. Liu, Y. Wang, and Y. Wang, "Automatic interactive evaluation for large language models with state aware patient simulator," *arXiv preprint arXiv:2403.08495*, 2024.

[24] R. Louie, A. Nandi, W. Fang, C. Chang, E. Brunskill, and D. Yang, "Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles," *arXiv preprint arXiv:2407.00870*, 2024.

[25] S. Chen, M. Wu, K. Q. Zhu, K. Lan, Z. Zhang, and L. Cui, "Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation," *arXiv preprint arXiv:2305.13614*, 2023.

[26] M. Gomes, J. Smith, and A. Patel, "Virtual patients and their impact on clinical communication skills training: A systematic review," *Journal of Medical Education Technology*, vol. 12, no. 4, pp. 245–259, 2023.

[27] O. Shaikh, V. E. Chai, M. Gelfand, D. Yang, and M. S. Bernstein, "Rehearsal: Simulating conflict to teach conflict resolution," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2024.

[28] D. Yang, C. Ziems, W. Held, O. Shaikh, M. S. Bernstein, and J. Mitchell, "Social skill training with large language models," *arXiv preprint arXiv:2404.04204*, 2024.

[29] E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, *et al.*, "Discovering language model behaviors with model-written evaluations," *arXiv preprint arXiv:2212.09251*, 2022.

[30] S. Lee, S. Lim, S. Han, G. Oh, H. Chae, J. Chung, M. Kim, B.-w. Kwak, Y. Lee, D. Lee, *et al.*, "Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics," *arXiv preprint arXiv:2406.14703*, 2024.

[31] N. B. Petrov, G. Serapio-García, and J. Rentfrow, "Limited ability of llms to simulate human psychological behaviours: a psychometric analysis," *arXiv preprint arXiv:2405.07248*, 2024.

[32] W. Mieleszczenko-Kowszewicz, D. Płudowski, F. Kołodziejczyk, J. Świstak, J. Sienkiewicz, and P. Biecek, "The dark patterns of personalized persuasion in large language models: Exposing persuasive linguistic features for big five personality traits in llms responses," *arXiv preprint arXiv:2411.06008*, 2024.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022.

[34] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 11, pp. 2755–2766, 2019.

[35] K. Gandhi, Z. Lynch, J.-P. Fränken, K. Patterson, S. Wambu, T. Gerstenberg, D. C. Ong, and N. D. Goodman, "Human-like affective cognition in foundation models," *arXiv preprint arXiv:2409.11733*, 2024.

[36] J. Gratch and S. Marsella, "A domain-independent framework for modeling emotion," *Cognitive Systems Research*, vol. 5, no. 4, pp. 269–306, 2004.

[37] J. D. Brett, R. Becerra, M. T. Maybery, and D. A. Preece, "The psychometric assessment of empathy: Development and validation of the perth empathy scale," *Assessment*, vol. 30, no. 4, pp. 1140–1156, 2023. PMID: 35435013.

[38] J. Broekens, D. DeGroot, and W. A. Kosters, "Formal models of appraisal: Theory, specification, and computational model," *Cognitive Systems Research*, vol. 9, no. 3, pp. 173–197, 2008.

[39] R. Krznaric, *Empathy: Why it matters, and how to get it*. TarcherPerigee, 2015.