PATIENT-Ψ: Using Large Language Models to Simulate Patients for Training Mental Health Professionals

Ruiyi Wang[*1], **Stephanie Milani**[*1], **Jamie C. Chiu**[2], **Jiayin Zhi**[1], **Shaun M. Eack**[3], **Travis Labrum**[3], **Samuel M. Murphy**[3], **Nev Jones**[3], **Kate Hardy**[4], **Hong Shen**[1], **Fei Fang**[1], **Zhiyu Zoey Chen**[1]

[1]School of Computer Science, Carnegie Mellon University,
[2]Department of Psychology, Princeton University,
[3]School of Social Work, University of Pittsburgh,
[4]Department of Psychiatry and Behavioral Sciences, Stanford University
{ruiyiwan,smilani}@andrew.cmu.edu, zhiyu.chen2@utdallas.edu

Abstract

Mental illness remains one of the most critical public health issues. Despite its importance, many mental health professionals highlight a disconnect between their training and actual real-world patient practice. To help bridge this gap, we propose PATIENT-Ψ, a novel patient simulation framework for cognitive behavior therapy (CBT) training. To build PATIENT-Ψ, we construct diverse patient cognitive models based on CBT principles and use large language models (LLMs) programmed with these cognitive models to act as a simulated therapy patient. We propose an interactive training scheme, PATIENT-Ψ-TRAINER, for mental health trainees to practice a key skill in CBT – formulating the cognitive model of the patient – through role-playing a therapy session with PATIENT-Ψ. To evaluate PATIENT-Ψ, we conducted a comprehensive user study of 13 mental health trainees and 20 experts. The results demonstrate that practice using PATIENT-Ψ-TRAINER enhances the perceived skill acquisition and confidence of the trainees beyond existing forms of training such as textbooks, videos, and role-play with non-patients. Based on the experts' perceptions, PATIENT-Ψ is perceived to be closer to real patient interactions than GPT-4, and PATIENT-Ψ-TRAINER holds strong promise to improve trainee competencies. Our code and data are released[1].

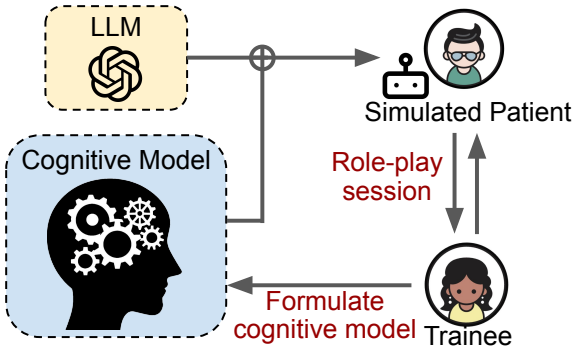


Figure 1: Illustration of our patient simulation idea.

1 Introduction

One in eight people globally are living with mental health conditions (World Health Organization, 2023)[2]. However, there is a significant gap between the available mental health support and patient needs, with over half (54.7%) of adults with a mental illness receiving no treatment in the US[3]. Training mental health professionals requires extensive effort, yet many professionals highlight a disconnect between their training and the complexities of real patient interactions. To understand these training challenges, we conducted a formative study involving semi-structured interviews with twelve mental health experts and trainees. This diverse group comprised of clinical psychologists, licensed social workers, and current master's students in social work. The experts provided insights into the difficulties faced when transitioning from formal CBT training to real-world practice (details in Appendix B). All experts noted that their training did not adequately prepare them for the unpredictable and multifaceted nature of real patient interactions. Despite wanting more interactive experiences, they found role-playing exercises with peers, a common training method, to be unrealistic, as these exercises often do not reflect actual therapy sessions.

There has been growing interest in developing LLM-based methods for psychology (Demszky et al., 2023; Chen et al., 2023b). In (Bubeck et al., 2023; Kosinski, 2023), ChatGPT and GPT-4 are able to solve some basic theory of mind tasks that generally require the ability to understand and attribute mental states to oneself and others. Inspired by such promise, we propose to use LLMs to simulate patients to train mental health professionals,

*Major contributors. See §A for individual contributions.
[1]https://github.com/ruiyiw/patient-psi
[2]https://www.who.int/campaigns/world-mental-health-day/2023
[3]https://mhanational.org/issues/2023/mental-health-america-access-care-data

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12772–12797
November 12-16, 2024

with the goal of bridging the gap between their existing training methods and the complexities of real patient interactions. However, two major challenges must be addressed to realize this idea:
Fidelity. *How can we build simulated patients that closely resemble the communicative behaviors of real patients with mental health disorders?*
Effectiveness. *How can we design an effective training scheme that allows trainees to benefit from interacting with these simulated patients?*

In this work, we claim that integrating a patient's *cognitive model* with an LLM can achieve high fidelity in simulating real patients with mental health disorders corresponding to that cognitive model. We implement this idea using the cognitive modeling framework in CBT (Beck, 2020), a popular paradigm in psychotherapy. We propose PATIENT-Ψ, a novel simulated patient agent that integrates cognitive modeling with LLMs. We collaborate with clinical psychologists to curate a dataset, PATIENT-Ψ-CM, which comprises 106 high-quality and diverse patient cognitive models. These cognitive models cover unhealthy cognitive structures embedded in multiple contexts, such as family issues, relationship problems, workplace challenges, and more. We then use these cognitive models to program an LLM to act as the PATIENT-Ψ agent. To better resemble the complex dynamics of real patient communications within a therapy session, we also integrate six conversational styles into PATIENT-Ψ. These conversational styles were identified from our formative study with mental health domain experts.

In CBT, formulating a patient's cognitive model is a crucial skill that therapists need to learn (Beck, 2020). Our design of PATIENT-Ψ naturally incorporates a feedback mechanism for trainees to practice this skill without extensive need for supervisor intervention, which is a desired benefit of AI-based training. We propose PATIENT-Ψ-TRAINER, an interactive training framework for mental health trainees to practice CBT cognitive model formulation using PATIENT-Ψ. Specifically, trainees converse with the simulated patient, PATIENT-Ψ, to formulate its cognitive model. Afterward, the system displays the original cognitive model that was used to program the simulated patient as a *reference*, allowing trainees to compare their results as feedback. Within this training framework, the effectiveness of the feedback theoretically depends on how accurately PATIENT-Ψ simulates a real patient with the corresponding cognitive model. Figure 1 illustrates the overall idea of our framework.

To evaluate the fidelity of PATIENT-Ψ and the effectiveness of PATIENT-Ψ-TRAINER, we conducted a thorough user study with **20 mental health experts** and **13 trainees**. Evaluation results from the experts indicate that: (1) PATIENT-Ψ closely resembles real patients in terms of maladaptive cognitions, conversational styles, and emotional states; outperforming GPT-4. (2) Practicing with PATIENT-Ψ-TRAINER is perceived to be highly beneficial for improving CBT formulation skills and better-preparing trainees for interactions with real patients. Experts also highlighted several advantages of PATIENT-Ψ-TRAINER, including customized options to choose conversation styles and the diverse patient cognitive models. Evaluation results from the trainees indicate that practicing with PATIENT-Ψ-TRAINER is perceived to improve skill and confidence, compared to current training methods. Overall, experts and trainees prefer using PATIENT-Ψ-TRAINER over a strong GPT-4 baseline. We also demonstrate that automatic evaluations with LLMs fail to assess the simulated patient fidelity, indicating the challenge of our task. Our contributions are summarized as follows:

- We propose PATIENT-Ψ, a novel simulated therapy patient, built using cognitive models grounded in psychology principles and LLMs.
- We propose PATIENT-Ψ-TRAINER, an interactive training framework for trainees to practice CBT formulation skills on PATIENT-Ψ.
- We create and release a dataset, PATIENT-Ψ-CM, with high-quality CBT-based cognitive models curated by clinical psychologists.
- Our user study with both mental health trainees and experts demonstrates that PATIENT-Ψ exhibits high fidelity to real patients, and practicing with PATIENT-Ψ-TRAINER significantly improves perceived skills and confidence in CBT formulation.

2 Methodology

In this section, we first describe the construction of PATIENT-Ψ in §2.1. We detail the integration of *cognitive models* with LLMs and the incorporation of *conversational styles* to accurately mimic real patient interactions. Next, we explain the training framework, PATIENT-Ψ-TRAINER, in §2.2, which utilizes PATIENT-Ψ to create an interactive learning environment for practicing CBT formulation skills. Figure 2 provides an overview of our method.

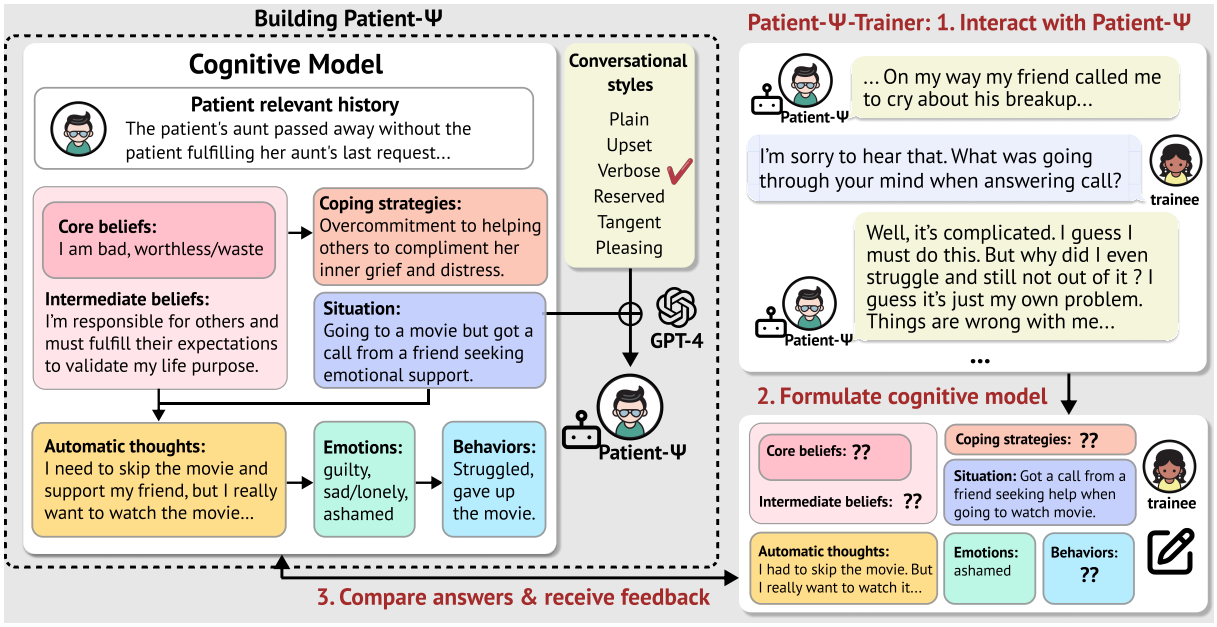


Figure 2: The overall framework of PATIENT-Ψ and PATIENT-Ψ-TRAINER. We integrate the expert-created cognitive model with GPT-4 to build PATIENT-Ψ. In PATIENT-Ψ-TRAINER, the trainee role-plays a therapy session with PATIENT-Ψ to formulate its cognitive model. The trainee can compare their formulation with the cognitive model used to build PATIENT-Ψ to get feedback.

2.1 PATIENT-Ψ

Using Cognitive Models to Simulate Patients. *Cognitive models* in mental health provide a structured framework for understanding how an individual's thoughts and beliefs are interconnected and influence emotions and behaviors. In established therapy paradigms like CBT (Beck, 2020), formulating a patient's cognitive model is central for a therapist to understand and address the maladaptive cognitions maintaining distress and symptoms (Hollon and Beck, 2013; Hofmann et al., 2012).

The *Cognitive Conceptualization Diagram* (CCD) (Beck, 2020) is a commonly used representation of a patient's cognitive model in CBT. The left side of Figure 2 depicts an example CCD-based cognitive model, illustrating eight key components. ① *Relevant history* contains significant past events that contribute to an individual's mental state. ② *Core Beliefs* are deeply ingrained perceptions about oneself, others, and the world. ③ *Intermediate beliefs* are the underlying rules, attitudes, and assumptions derived from core beliefs and shape an individual's thought patterns. ④ *Coping strategies* are techniques used to manage negative emotions. An external event or context (⑤ *a situation*) may trigger quick, evaluative thoughts without deliberation (⑥ *automatic thoughts*) stemming from the beliefs, leading to responses in terms of ⑦ *emotions* and ⑧ *behaviors*. A CCD-based cognitive model links the components together, creating a framework for identifying and understanding patients' underlying cognitive processes. For all the components, we adopt the definitions and formulations put forth by (Beck, 2020). These include: three major core beliefs (②)—helpless, unlovable, and worthless—each with several fine-grained core beliefs, for a total of 19 core belief categories; 9 emotion (⑦) categories; the rest of the components are formulated as free-text entries. See Table 2 and Appendix C.1 for the categories. In this work, we integrate CCD-based cognitive models into an LLM to simulate patients whose communication reflects the underlying cognitive processes.

The PATIENT-Ψ-CM Cognitive Model Dataset. To the best of our knowledge, no existing work offers a dataset of realistic cognitive models due to two challenges: 1) the data privacy constraints involved in acquiring real patient cognitive models and 2) the high-level expertise required to perform manual creations. In this work, we propose the first dataset of CCD-based cognitive models grounded in CBT principles, PATIENT-Ψ-CM, created by clinical psychologists. We first prompt GPT-4 Turbo (OpenAI, 2023) to create summaries from therapy session transcripts. These transcripts were obtained from the Alexander Street database[4] under the subject "Counseling and Therapy" and

[4] https://alexanderstreet.com/, accessed through our institution's subscription.

Style	Description
`plain`	Direct, straightforward.
`upset`	Frustration, resistance.
`verbose`	Overly expressive.
`reserved`	Minimal, restrained.
`tangent`	Deviates from the main topic.
`pleasing`	Agreeable to a fault.

Table 1: Different conversational styles that PATIENT-Ψ can take on, with descriptions. More detailed examples in Appendix C.3. Yellow styles are harder; blue style is easier.

Situations	#	**Emotions**	#
family dynamics	25	anxious	60
workplace pressure	20	sad	50
relationship dynamics	19	angry	22
social interactions	18	hurt	21
personal growth issues	8	disappointed	19
financial concerns	8	ashamed	17
daily life stressors	8	guilty	13
		suspicious	2
Core beliefs	#	jealous	1
helpless	94		
unlovable	71		
worthless	15	**106 cognitive models**	

Table 2: PATIENT-Ψ-CM statistics. Details in Appendix C.1.

the keyword "Cognitive Behavioral Therapy". Two clinical psychologists then manually create cognitive models by drawing inspiration from the transcript summaries, incorporating their professional expertise, and applying their creativity (within clinical constraints). This process involves developing new cases inspired by the summaries and composing the corresponding cognitive models. We emphasize *diversity* and *realism* to the psychologists when creating the models. We end up with a dataset containing 106 cognitive models (an example is shown in Figure 2, left). Each cognitive model is associated with one activating situation. See Appendix C.2 for details of dataset creation and more example cognitive models from PATIENT-Ψ-CM.

Conversational Styles Integration. In the formative study (Appendix B), domain experts emphasized that real patients exhibit different *conversational styles* during therapy. Based on these discussions, we identify six styles for PATIENT-Ψ, detailed in Table 1. To create a natural curriculum, the styles are two levels of difficulty. The easiest style, `plain`, features direct and straightforward communication. The more difficult styles require trainees to exert more effort to elicit relevant information. To incorporate these styles with PATIENT-Ψ, two clinical psychologists annotate each cognitive model with a list of valid conversational styles and develop instructions for PATIENT-Ψ to simulate a patient for each style. Detailed descriptions and examples of the conversational styles are provided in Appendix C.3.

Patient Agent Simulation. We prompt GPT-4 to build PATIENT-Ψ which consists of a patient's cognitive model, the conversational style prompt, and a list of instruction prompts. Appendix C.4 contains the full prompts. The model is continually prompted to engage in a CBT-based therapy session, role-playing a patient with the corresponding cognitive model and conversational styles.

2.2 PATIENT-Ψ-TRAINER

With the development of PATIENT-Ψ, we introduce PATIENT-Ψ-TRAINER, an interactive training framework designed for mental health professionals to practice cognitive model formulation for CBT. PATIENT-Ψ-TRAINER offers a structured, three-step training process: 1) engaging with PATIENT-Ψ in a simulated CBT session, 2) formulating PATIENT-Ψ's cognitive model through interaction, and 3) reviewing the original cognitive model used to create PATIENT-Ψ for feedback. The right-hand side of Figure 2 illustrates this process.

Training Process. Trainees begin by choosing one of the six conversational styles from PATIENT-Ψ-TRAINER's web application interface (screenshots in Appendix G). Then, a patient is generated using the chosen style and a randomly-selected cognitive model from PATIENT-Ψ-CM compatible with that style. The interface displays the patient's relevant history in preparation for the session. During this session, the trainee engages with PATIENT-Ψ, applying their therapeutic skills with the goal of formulating the CCD-based cognitive model used to program PATIENT-Ψ. This involves eliciting and summarizing all cognitive elements underlying the conversation with PATIENT-Ψ.

Real-Time Feedback. Upon concluding the interactive session, PATIENT-Ψ-TRAINER allows the trainee to compare their formulated cognitive model with the original cognitive model used to program PATIENT-Ψ. This side-by-side comparison highlights discrepancies, providing detailed feedback. Trainees can continue to chat with PATIENT-Ψ to refine their formulations. This natural feedback loop, stemming from our design of using the cognitive model to program the patient,

offers the advantage of minimal human supervision efforts, enabling trainees to practice independently.

3 Experiment Setup

We now present the experimental setup for evaluating PATIENT-Ψ and PATIENT-Ψ-TRAINER. We aim to answer the following research questions:

RQ 1 **Fidelity:** Does PATIENT-Ψ improve the fidelity of patient simulations compared to baselines?

RQ 2 **Accuracy:** How closely does PATIENT-Ψ imitate the underlying cognitive model?

RQ 3 **Effectiveness:** Do experts and trainees perceive PATIENT-Ψ-TRAINER as an effective tool for CBT training?

RQ 4 **AutoEval:** Can we leverage existing methods, such as LLMs, to automatically evaluate the patient simulations?

In §4, we answer the first three RQs through our user study with both trainees and experts. Then, in §5, we show that current automatic evaluations cannot capture the nuances necessary for conducting highly technical, domain-specific assessments. This finding not only shows the importance of user study evaluations but also motivates future work on performant automatic evaluators.

Evaluation Dimensions. We design a set of fine-grained dimensions to assess each RQ, using insights from the formative study and existing literature (Beck, 2020; Bouter et al., 2012; Issenberg et al., 2005; Silverman et al., 2013; Ekman, 1992). To ensure that the simulated patients' responses reflect those of real patients, we measure the **fidelity** of the *emotional states*, *conversational styles*, and *maladaptive cognitions* of PATIENT-Ψ to real patients. To assess the **accuracy** of PATIENT-Ψ in emulating the underlying expert-validated cognitive model, we evaluate each component's accuracy. To assess the **effectiveness** of PATIENT-Ψ-TRAINER, we measure the perceived improvements of CBT formulation skills: *identifying maladaptive thinking patterns* and *identifying beliefs*. We also measure the perceived *confidence improvement* of the trainees. Finally, we assess *usability* to ensure the tool's ease of use for users. Due to space constraints, the usability results are in Appendix E.4.

For pairwise comparisons, the options are: "A is much better than B," "A is somewhat better than B," "about the same," "B is somewhat better than A," and "B is much better than A." We map the results to a scale from -2 to 2, where ± 2 indicates a strong preference. Individual measures use a 5-point Likert scale from 1 to 5, where 5 means "strongly agree" or "extremely accurate," and 1 means "strongly disagree" or "not accurate at all." Specific values for each dimension are in §4.

Baselines. We leverage vanilla GPT-4 with a general description of patients with depression or anxiety as the input, rather than the cognitive models (see Appendix D). Thus, we cannot show the reference cognitive model as feedback and do not include the conversational styles. We also compare with existing training techniques, which includes peer role-plays or textbook examples.

User Study Details. Assessing simulated therapeutic dialogue is a cognitively difficult process that requires professional training and experience, making typical crowdsourcing data collection approaches difficult. To ensure high-quality evaluations from those with significant real patient experience (experts) and from the population who would use PATIENT-Ψ-TRAINER in practice (trainees), we recruit 20 current mental health practitioners and 13 social work students, respectively.[5] §7 details the IRB approval and recruitment. Each participant practices with PATIENT-Ψ-TRAINER and the baseline in a randomized order, completing two simulated patient sessions for each. To ensure comprehensive evaluation across diverse cognitive models, we assign each participant simulated patients with distinct underlying cognitive models, covering a total of 66 cognitive models from PATIENT-Ψ-CM. For expert evaluations, we distribute two specific conversational styles to each participant to achieve an overall balanced distributions of all styles. Trainees can select two styles based on their expertise level and confidence. More protocol details are in Appendix D.

Automatic Evaluation Details To leverage LLMs' capabilities as a judge for evaluating open-ended tasks, we use state-of-art LLMs, including GPT-4 (OpenAI, 2023) and Llama 3 70B (AI@Meta, 2024), as evaluators. We set the temperature to 1.0 for both LLMs. For assessing fidelity, we prompt LLM evaluators with the same instructions used in the questionnaire and the con-

[5]We recruited participants through the professional networks of our co-authors in mental health (clinical psychologists and professors in clinical psychology and social work), as well as snowball sampling.

versations between experts and PATIENT-Ψ, as collected during the user study. The Likert scale ratings are mapped to numerical values ranging from 1 to 5 to present the results in a direct, interpretable format. For evaluating accuracy, we apply two approaches based on the types of the CBT components being assessed:

1. **Text-based fields** (including situation, coping strategies, intermediate beliefs, automatic thoughts, and behaviors): we generate a multiple-choice question by randomly sampling four additional components from PATIENT-Ψ-CM, alongside the ground-truth component. GPT-4 is then prompted to select the component most closely reflected in PATIENT-Ψ's conversations. We report the accuracy based on a 5-class classification task.

2. **Categorization fields** (including core beliefs and emotions): These components are selected from predefined sets of core beliefs and emotions as described in CBT literature (Beck, 2020). GPT-4 is prompted to choose the options that are reflected in the patient's conversations, and we report the F1 score for these selections.

4 User Study Results

4.1 RQ 1: Fidelity to Real Patients

To assess the fidelity of PATIENT-Ψ to real patients, experts compare existing training techniques, the GPT-4 baseline, and PATIENT-Ψ (Table 3). We ask experts for their overall impressions of these training methods, resulting in 20 data points for each comparison in this subsection. Paired t-tests show that PATIENT-Ψ significantly outperforms the other methods ($p < 10^{-4}$), indicating that PATIENT-Ψ provides the most realistic patients, addressing RQ 1 positively. This is promising for PATIENT-Ψ: our formative study highlighted a gap in trainee preparation for real interactions, which PATIENT-Ψ can effectively fill.

PATIENT-Ψ exhibits higher fidelity to real patients than the GPT-4 baseline. Each expert compares the fidelity dimensions (*emotional states*, *conversational styles*, *maladaptive cognitions*) of PATIENT-Ψ and the GPT-4 baseline to real patients. Figure 3 (left) depicts the distribution of expert comparisons; summary statistics in Table 11, Appendix E. PATIENT-Ψ is rated higher along all dimensions for fidelity: it better represents the maladaptive cognitions (μ: 0.6, $p < 0.05$), the emotional states (μ: 1.1, $p < 10^{-4}$), and the conversational styles (μ: 1.3, $p < 10^{-4}$) of real patients. Experts expressed that PATIENT-Ψ offered a more realistic challenge of extracting information from patients, unlike the baseline which was too forthcoming with responses. One expert noted that sessions with the baseline felt "almost like doing therapy with a therapist," highlighting the challenge of simulating real patient behavior — even with advanced LLMs likely pretrained on an extensive corpus of therapy knowledge.

Comparison	Patient Fidelity μ
PATIENT-Ψ vs. GPT-4	1.3***
PATIENT-Ψ vs. Traditional	1.3***
GPT-4 vs. Traditional	0.7*
* : $p < 0.05$, ** : $p < 0.01$, *** : $p < 10^{-4}$	

Table 3: PATIENT-Ψ provides significantly more realistic simulated patients compared to the GPT-4 baseline and traditional methods. Closer to 2/-2: the first/second method is better.

4.2 RQ 2: Accuracy to Cognitive Model

To be practically useful, PATIENT-Ψ must accurately reflect the reference cognitive model, as trainees rely on it for feedback on their completed formulations. Experts evaluate PATIENT-Ψ's overall accuracy and its accuracy for each component of the cognitive model, resulting in 40 data points per dimension. Table 12, Appendix E.2 presents the summary statistics; Figure 4 illustrates the distribution. The results are promising: overall, PATIENT-Ψ is rated on average as *very accurate*. For each of the 8 components, PATIENT-Ψ is rated on average as *very* to *extremely* accurate. Specifically, 80-88% of the simulated patients achieve *very* to *extremely accurate* ratings for each of the 8 components, answering RQ 2. Crucially, since the reference cognitive model is accurately captured by PATIENT-Ψ, trainees can rely on it to receive high-quality feedback on their responses.

Patient simulation may involve an accuracy-fidelity trade-off. Expert feedback from our study reveals insights into the challenges of accurately simulating patients in alignment with the underlying cognitive model. Specifically, there exists a tension between some of the evaluation metrics. For example, one expert, noting the limitations of text-only interfaces, suggests increasing the use

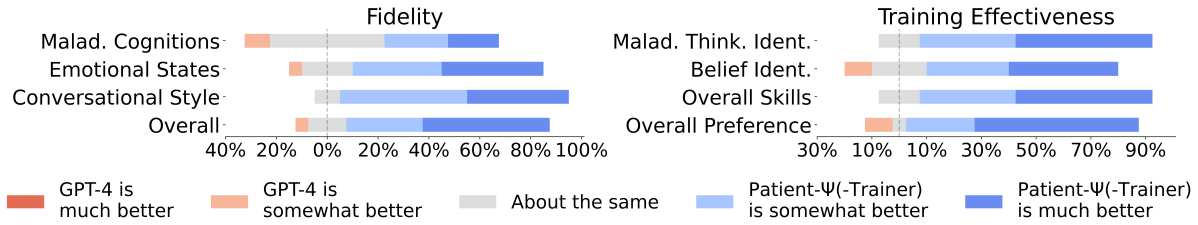


Figure 3: Fidelity of PATIENT-Ψ and training effectiveness of PATIENT-Ψ-TRAINER compared to GPT-4 baseline along multiple dimensions. X-axis: the % of experts who voted for a specific option; y-axis: the assessment dimension. Malad. means maladaptive, Think. means thinking, and Ident. means identification. PATIENT-Ψ more closely resembles real patients (fidelity, left) and is considered more effective for trainees (training effectiveness, right).

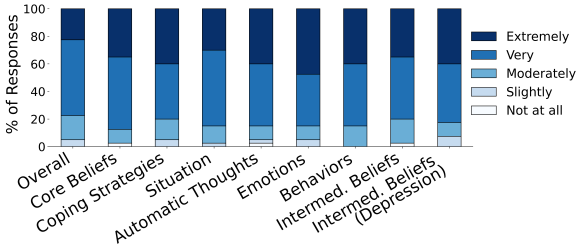


Figure 4: Experts rate 97% of the PATIENT-Ψ patients as at least moderately accurate in reflecting the reference cognitive model. Intermed. means Intermediate.

Comparison	Effectiveness μ	
	Expert	Trainee
PATIENT-Ψ-TRAINER vs. GPT-4	1.4***	1.1**
PATIENT-Ψ-TRAINER vs. Traditional	1.7***	1.6***
GPT-4 vs. Traditional	1.2***	1.0**

$*: p < 0.05$, $**: p < 0.01$, $***: p < 10^{-4}$

Table 4: Experts and trainees find PATIENT-Ψ-TRAINER to be significantly more effective for improving overall skills compared to the GPT-4 baseline and traditional methods. Closer to 2/-2: the first/second method is better.

of explicit emotion words to improve the model's ability to accurately convey emotions. However, this approach potentially conflicts with real-world language patterns, as highlighted by another expert. This expert works with populations from prisons. According to the expert, this population tends to not use any feeling words, so the expert believes that including emotion words is less realistic based on their experience with this subpopulation. As a result, including such words may improve accuracy but could do so at the cost of fidelity. This finding suggests that, in some cases, there exists a tension between evaluation metrics.

4.3 RQ 3: Effectiveness for Training

Experts and trainees provide their perception of the effectiveness of PATIENT-Ψ-TRAINER and the GPT-4 baseline compared to existing training techniques. In this section, we have 20 comparison points for the experts and 13 for the trainees, as we ask them to provide their *overall* assessment of the tool, not individual patients. Paired t-tests reveal that experts and trainees perceive PATIENT-Ψ-TRAINER as significantly more effective at improving overall skills than both traditional techniques ($p < 10^{-4}$) and the GPT-4 baseline ($p < 0.01$) (Table 4), answering RQ 3. Compared to trainees with limited real patient experience, experts show stronger preferences for our system, further demonstrating PATIENT-Ψ-TRAINER's effectiveness in preparing for real patient interactions. Compared to traditional methods *without* real patient interactions, experts favor PATIENT-Ψ-TRAINER's ease of access (90%), customization options of different conversational styles (90%), and interactive experience (65%). Compared to practicing *with* real patients, experts value PATIENT-Ψ-TRAINER's ease of access (79%), customization options of different conversational styles (88%), and safer setting for training (88%). After only two sessions with our tool, one trainee remarked that it "helped to make things more clear with the CCD (cognitive model), for my training/class it was somewhat meaningless and challenging to build one."

PATIENT-Ψ-TRAINER is a more effective training tool than the GPT-4 baseline. Both groups compare PATIENT-Ψ-TRAINER and the GPT-4 baseline along the fine-grained dimensions. Figure 3 (right) shows the distribution of expert comparisons; summary statistics for both groups in Table 13, Appendix E. Both groups indicate that PATIENT-Ψ-TRAINER would be significantly more effective at improving the key CBT skills of identifying beliefs (μ: 1.0, $p < 0.01$; μ: 0.9, $p < 0.05$, respectively) and maladaptive thinking (μ: 1.4, $p < 10^{-4}$; μ: 1.0, $p < 0.01$, respectively). Furthermore, both groups overwhelmingly prefer PATIENT-Ψ-TRAINER for practical use (both μ: 1.4, $p < 10^{-4}$), showing its high potential for real-

Comparison	Confidence Improvement μ
PATIENT-Ψ vs. GPT-4	1.2**
PATIENT-Ψ vs. Traditional	1.8***
GPT-4 vs. Traditional	1.4***

$^*: p < 0.05$, $^{**}: p < 0.01$, $^{***}: p < 10^{-4}$

Table 5: PATIENT-Ψ can provide significantly greater confidence improvement vs. the GPT-4 baseline and traditional methods. Closer to 2/-2: the first/second method is better.

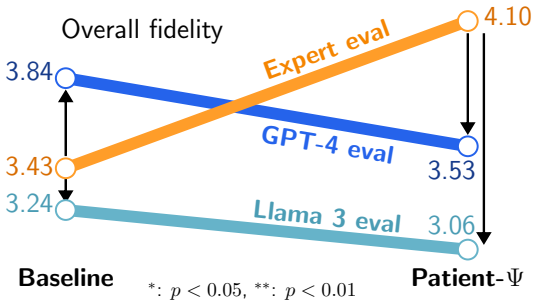


Figure 5: Mean overall fidelity of PATIENT-Ψ and baseline as evaluated by experts and LLMs. Compared to experts, both GPT-4 and Llama 3 demonstrate opposite trends.

Text-based	Acc.	Categorization	F1	F1 (expert)
Situation	0.97	Core beliefs	0.80	0.77
Coping strategies	0.93	Emotions	0.72	0.74
Intermediate beliefs	0.92	Core beliefs (fine-grained)	0.48	0.44
Automatic thoughts	0.88			
Behaviors	0.84			

Table 6: Accuracy and Macro F1 of PATIENT-Ψ evaluated by GPT-4. For text-based fields, GPT-4 is prompted to select the components among four distractor options randomly sampled from PATIENT-Ψ-CM. For categorization, GPT-4 is prompted to select all relevant categories of emotions and core beliefs.

world impact.

PATIENT-Ψ-TRAINER can improve trainees' confidence over the baseline and traditional methods. Toward our aim of improving preparation for real patient interactions, trainees compare their perceived confidence improvement when using PATIENT-Ψ-TRAINER versus traditional methods and the GPT-4 baseline. They rate PATIENT-Ψ-TRAINER as significantly more effective at boosting their confidence (Table 5).

Experts unanimously find value in PATIENT-Ψ-TRAINER 's real-time feedback. A core feature of PATIENT-Ψ-TRAINER is the real-time feedback provided by displaying the accurate reference cognitive model (§4.2). 100% of experts prefer that PATIENT-Ψ-TRAINER display the reference cognitive model at the end of training and unanimously agree that viewing it is beneficial for practicing CBT skills. One expert emphasized, "Without the answers, I think it's much less helpful."

Experts unanimously prefer PATIENT-Ψ-TRAINER's option to practice with different conversational styles. Another core feature of our method is the option to practice with patients exhibiting different conversational styles. 100% of experts prefer this option. One expert noted that the styles "are more reflective of actual patients" and can be linked to specific diagnoses and symptoms, making the interactions more accurate. Nearly all experts (95%) view this feature as useful for interacting with diverse real patients and improving trainee confidence for real interactions. These results suggest that offering diverse patient types is critical for effective and realistic training.

5 Automatic Evaluation Results

Given the potential of using LLMs for evaluating text generation quality (Chiang and Lee, 2023), we attempt to automatically assess **the fidelity and accuracy of PATIENT-Ψ and the baseline** using two state-of-the-art LLMs as evaluators: GPT-4 (OpenAI, 2023) and Llama 3 70B (AI@Meta, 2024). We evaluate over the 40 conversation histories between the experts and PATIENT-Ψ in our user study.

LLM-based evaluators tend to underestimate PATIENT-Ψ's fidelity in favor of GPT-4 baseline. Following RQ 1 (Fidelity), the LLMs are prompted to provide ratings on a 5-point Likert scale assessing the fidelity of how closely the simulated patient resembles real patients following the same dimensions used in the user study. In Figure 5, paired t-tests show that the fidelity of PATIENT-Ψ, as evaluated by both LLMs (μ: 3.53; μ: 3.06, respectively), is consistently lower than expert evaluation (μ: 4.10; $p < 0.01$ for the differences between expert and both LLMs), contrasting with the user study results. GPT-4 assigns the highest fidelity scores to the GPT-4 baseline. All fidelity dimensions demonstrate the same trend (see Appendix F).

GPT-4 assesses PATIENT-Ψ's accuracy similarly to experts. To evaluate the accuracy of PATIENT-Ψ in reflecting the underlying cognitive models, we design proxy measures to prompt GPT-4 to select the closest cognitive model components reflected by the conversation. As shown in Table 6, GPT-4 achieves high accuracy in most components, except for fine-grained core beliefs, where there are 19 categories and demonstrate high variance by nature.

GPT-4 achieves similar scores with the experts' inputs, suggesting the high accuracy of PATIENT-Ψ in representing the underlying cognitive models, aligning with the experts' evaluations.

The results suggest that GPT-4 excels in understanding cognitive models from patients' conversations, attributable to its extensive acquisition of CBT knowledge during pre-training. However, it falls short in assessing the realism of patients. This aligns with our findings that the GPT-4 baseline fails to create high-fidelity patient simulations. While it accurately conveys CBT knowledge, it does so in a manner resembling a therapist speaking directly and explicitly, rather than a real patient whose conversation naturally reflects their disorders. This underscores the challenges and contributions of our work, highlighting the difficulty of crafting realistic patient interactions even with the most powerful LLMs today.

6 Related Work

Our work is broadly related to the recent use of LLMs in psychology, education, and computational social science (Hsu et al., 2023; Chiu et al., 2024; Fu et al., 2023; Ji et al., 2022; Zanwar et al., 2023; Juhng et al., 2023; Ziems et al., 2024; Halder et al., 2017; Sharma et al., 2020b,a; Atapattu et al., 2022; Mishra et al., 2023; Sonkar et al., 2023; Wang et al., 2024; Zhou et al., 2024). In contrast to existing research on using LLMs for CBT, which focuses on cognitive distortion detection (Shreevastava and Foltz, 2021; Ding et al., 2022; Lybarger et al., 2022; Chen et al., 2023b) and negative thoughts reframing (Sharma et al., 2023, 2024), our work aims to provide realistic and interactive scenarios for CBT professional development by simulating diverse patient types using LLMs. As a result, our work most closely relates to research that leverages LLMs for simulation-based training, particularly communication skill learning and emotion management grounded in dialectical behavioral therapy (Lin et al., 2024), social skill training (Yang et al., 2024), and clinical diagnosis (Chen et al., 2023a). Our work is the first to ground LLM-based simulations in clinical psychology theory by leveraging CBT-based cogntive models to program LLMs, incorporate a natural curriculum and feedback mechanism in the training tool, and perform evaluation in context with mental health trainees and professionals rather than crowdworkers.

7 Conclusion

In this paper, we introduce PATIENT-Ψ, a simulated patient that integrates cognitive models with an LLM to accurately mimic the communicative behaviors of real patients. We propose PATIENT-Ψ-TRAINER, where trainees engage in role-playing therapy sessions with PATIENT-Ψ and attempt to formulate the underlying cognitive model. User studies with both mental health experts and trainees demonstrate the high fidelity of PATIENT-Ψ and the training effectiveness of PATIENT-Ψ-TRAINER, showing improvements over existing training methods and outperforming a GPT-4 baseline. Our framework has the potential to transform mental health professional training and be generalized to broader training protocols and therapy paradigms.

Limitations

In this work, we evaluate our framework using GPT-4. As we do not rely on specific properties of GPT-4, we believe the framework could be applied to any powerful open-source LLMs such as Llama 3 (Dubey et al., 2024) and Gemma (Team et al., 2024). For future work, it would be interesting to evaluate various generative models and prompting techniques on this task. Additionally, in this work, our measures of the training effectiveness are all perceived improvements from the participants after they practice with PATIENT-Ψ-TRAINER for two sessions. Measuring objective skill improvements could take the form of longitudinal randomized controlled trials (RCTs). Conducting these RCTs would also help address another limitation of our study, the sample size. Due to how specialized the participants must be to properly evaluate the tools and the $1-2$ hours required to conduct each user study, the sample size of our study is only 33 in total. Our results are statistically significant; however, RCTs would enable us to study the tools with a larger population. We leave this for future work. Finally, while we primarily target CBT cognitive formulation training in this paper, we believe our methodology can be generalised to other training protocols and therapy paradigms.

Ethics Statement

IRB (Institutional Review Board) Approval. This project is approved by our Institutional Review Board (IRB) with study number `STUDY2023_00000451`. For the creation of cognitive models, any other annotation work, as well as consultations, we collaborate with clinical psychologists and professors in clinical psychology and social work, who are our co-authors. For both the formative study and user study, we recruited participants through the professional networks of our co-authors, as well as snowball sampling. Experts are defined as those with a graduate degree in clinical psychology, social work, or other related majors and have worked with at least 5 patients. Trainees are those still in school/training or with fewer than 5 real patient experiences. For the formative study, we recruited a total of 12 participants. We pay a $30 Amazon gift card for each participant for a 30-minute session over Zoom. For the user study, we recruited a total of 33 participants. We pay a $60 Amazon gift card for a 60-90-minute session over Zoom.

Informed Consent. All participants in the user study and formative study were 18 or older and provided informed consent. We did not assess any clinical outcomes. All data collected from the participants were de-identified and consented to be released for research purposes.

Crisis Resources The risk to the participants is minimal, no greater than their professional working or training environment of mental health support in the context of conducting therapy sessions with people with mental health issues. Nevertheless, we do not exclude the possibility that some AI-generated content might still be upsetting to the participants. Therefore, we advise participants to use a free crisis resource available at `https://www.7cups.com/` if needed, and they are free to terminate the study at any time without facing any negative consequences. This risk assessment and crisis resource information have been included in our IRB approval and provided as part of the informed consent to participants.

System and Data Usages. All the data and systems developed in this work are intended solely for academic research purposes. The systems developed in this work are intended to augment existing mental health training, not to replace it. One major benefit of our system, as highlighted by experts in the user study, is that it provides trainees with a safe training environment. By working with AI patients, trainees can practice without the risk of causing actual harm due to mistakes made during simulated therapy sessions. Our system is designed for academic and educational purposes only. Real-world deployments will require further work, including measuring objective skill improvements and developing protocols for integrating the system with existing training methods, all within the framework of large-scale randomized controlled trials (RCTs).

We utilize therapy session transcripts from the Alexander Street database[6], accessed through our institution subscription. Our usage complies with their fair use policy. GPT-4 is employed to generate summaries of these transcripts. For constructing the cognitive model dataset, two clinical psychologists manually create cognitive models based on inspirations from the transcript summaries, clinical experience, and creativity—effectively generating new cases. The resulting dataset is manually verified and does not contain any Personally Identi-

[6] `https://alexanderstreet.com/`

fiable Information (PII). It is intended solely for academic research purposes and will be made available only to academic institutions with subscriptions to the Alexander Street database. The dataset will be released upon request.

Acknowledgments

This work is supported in part by NSF grant IIS-2046640 (CAREER) and by Sloan Research Fellowship. Hong Shen is supported by an award from the Public Interest Technology University Network Challenge (PIT-UN) and an award from the Carnegie Mellon University Block Center for Technology and Society (Award No. 59201.1.5007718).

References

AI@Meta. 2024. Llama 3 model card.

Thushari Atapattu, Mahen Herath, Charitha Elvitigala, Piyanjali de Zoysa, Kasun Gunawardana, Menasha Thilakaratne, Kasun de Zoysa, and Katrina Falkner. 2022. EmoMent: An emotion annotated mental health corpus from two South Asian countries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6991–7001, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Judith S Beck. 2020. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications.

Shifra Bouter, Evelyn Weel-Baumgarten, and Sanneke Bolhuis. 2012. Construction and validation of the nijmegen evaluation of the simulated patient (nesp): Assessing simulated patients' ability to role-play and provide feedback to students. *Academic medicine : journal of the Association of American Medical Colleges*, 88.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Siyuan Chen, Mengyue Wu, Kenny Q. Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023a. Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation. *Preprint*, arXiv:2305.13614.

Zhiyu Chen, Yujie Lu, and William Wang. 2023b. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *Preprint*, arXiv:2401.00820.

Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel Jones-Mitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2:688–701.

Xiruo Ding, Kevin Lybarger, Justin Tauscher, and Trevor Cohen. 2022. Improving classification of infrequent cognitive distortions: Domain-specific model vs. data augmentation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 68–75, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Guanghui Fu, Qing Zhao, Jianqiang Li, Dan Luo, Changwei Song, Wei Zhai, Shuo Liu, Fan Wang, Yan Wang, Lijuan Cheng, Juan Zhang, and Bing Xiang Yang. 2023. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. *Preprint*, arXiv:2308.15192.

Kishaloy Halder, Lahari Poddar, and Min-Yen Kan. 2017. Modeling temporal progression of emotional status in mental health forum: A recurrent neural net approach. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 127–135, Copenhagen, Denmark. Association for Computational Linguistics.

Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*, 36:427–440.

Steven D Hollon and Aaron T Beck. 2013. Cognitive and cognitive-behavioral therapies. *Bergin and Garfield's handbook of psychotherapy and behavior change*, 6:393–442.

Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2023. Helping the helper: Supporting peer counselors via ai-empowered practice and feedback. *arXiv preprint arXiv:2305.08982*.

Barry Issenberg, William Mcgaghie, Emil Petrusa, David Gordon, and Ross Scalese. 2005. Features and uses of high-fidelity medical simulations that lead to effective learning: A beme systematic review*. *Medical teacher*, 27:10–28.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Swanie Juhng, Matthew Matero, Vasudha Varadarajan, Johannes Eichstaedt, Adithya V Ganesan, and H Andrew Schwartz. 2023. Discourse-level representations can improve prediction of degree of anxiety. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1500–1511.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

James R Lewis. 2018. The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction*, 34(7):577–590.

Inna Wanyin Lin, Ashish Sharma, Christopher Michael Rytting, Adam S Miner, Jina Suh, and Tim Althoff. 2024. Imbue: Improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction. *arXiv preprint arXiv:2402.12556*.

Kevin Lybarger, Justin Tauscher, Xiruo Ding, Dror Ben-Zeev, and Trevor Cohen. 2022. Identifying distorted thinking in patient-therapist text message exchanges by leveraging dynamic multi-turn context. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 126–136.

Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023. PAL to lend a helping hand: Towards building an emotion adaptive polite and empathetic counseling conversational agent. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12254–12271, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 2020a. Engagement patterns of peer-to-peer interactions on mental health platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 614–625.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020b. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–29.

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G Lucas, Adam S Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive reframing of negative thoughts through human-language model interaction. *arXiv preprint arXiv:2305.02466*.

Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 151–158.

Jonathan Silverman, Suzanne Kurtz, and Juliet Draper. 2013. *Skills For Communicating With Patients*.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. Class: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Graham Neubig, Yonatan Bisk, and Hao Zhu. 2024. Sotopia-π: Interactive learning of socially intelligent language agents. *Preprint*, arXiv:2403.08715.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*.

Sourabh Zanwar, Xiaofei Li, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2023. What to fuse and how to fuse: Exploring emotion and personality fusion strategies for explainable mental disorder detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8926–8940, Toronto, Canada. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. *Preprint*, arXiv:2310.11667.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Detailed Individual Contributions

All authors contribute to paper writing.
Ruiyi Wang: Co-lead, Model development, Interface construction, Data, User study.
Stephanie Milani: Co-lead, User study, Formative study.
Jiayin Zhi: Interface construction.
Jamie C. Chiu, Kate Hardy: Data creation, Consultations.
Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones: Consultations, User study.
Hong Shen, Fei Fang: Co-advising.
Zhiyu Zoey Chen: Overall project lead.

B Formative Study Details

To understand the challenges faced during CBT training and elicit feedback on a prototype of PATIENT-Ψ-TRAINER, we first conducted a formative study in the form of semi-structured interviews with trainees and experts in mental health.[7] This study was conducted over Zoom.

Participant Information. We interviewed twelve individuals who had diverse educational backgrounds and career experiences. Among them, five were Master's students, the rest included a Ph.D. student, a post-doctoral fellow, three licensed social workers, and two psychologists. Our participants also had varied levels of experience working with patients. Only one individual had not yet worked with any patients, while another reported working with anywhere from 1500-3000 patients over their career. We refer to individuals as *experts* if they received a graduate degree and have worked with at least 5 patients; we use *trainees* if they do not have a graduate degree and have formal experience with fewer than 5 patients. This definition is consistent with our user study. Thus, for our formative interviews, we have 5 trainees and 7 experts.

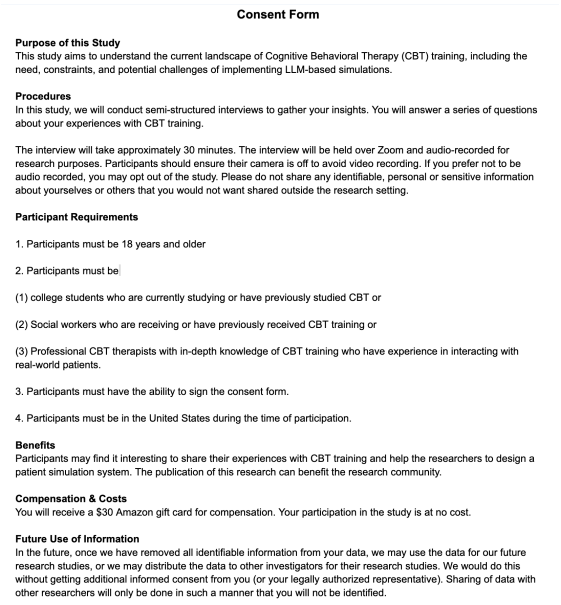
Instructions to Participants. Before each interview, the participant voluntarily signs the consent form. We provide the screenshots of the consent form with all sensitive information removed in Figures 6 and 7. After receiving the signed consent form, we then proceed with the interview. When the session starts, we remind participants of the recorded nature of the conversation and verbally summarize the goal of the interview. We also provide a high-level overview of the structure of the interview. We confirm consent to audio record the interview before proceeding. In our interviews, we first ask the experts questions about challenges they faced transitioning from their formal CBT training to practice. We then present both groups with a prototype of PATIENT-Ψ-TRAINER to elicit feedback.

B.1 Insights

We now elaborate on the main insights that we gleaned from this formative study.

Insight 1: Experts feel that their training did not adequately prepare them for real-world prac-

[7] We recruited participants through the professional networks of our co-authors in mental health (clinical psychologists and professors in clinical psychology and social work).



Consent Form

Purpose of this Study
This study aims to understand the current landscape of Cognitive Behavioral Therapy (CBT) training, including the need, constraints, and potential challenges of implementing LLM-based simulations.

Procedures
In this study, we will conduct semi-structured interviews to gather your insights. You will answer a series of questions about your experiences with CBT training.

The interview will take approximately 30 minutes. The interview will be held over Zoom and audio-recorded for research purposes. Participants should ensure their camera is off to avoid video recording. If you prefer not to be audio recorded, you may opt out of the study. Please do not share any identifiable, personal or sensitive information about yourselves or others that you would not want shared outside the research setting.

Participant Requirements

1. Participants must be 18 years and older

2. Participants must be

(1) college students who are currently studying or have previously studied CBT or

(2) Social workers who are receiving or have previously received CBT training or

(3) Professional CBT therapists with in-depth knowledge of CBT training who have experience in interacting with real-world patients.

3. Participants must have the ability to sign the consent form.

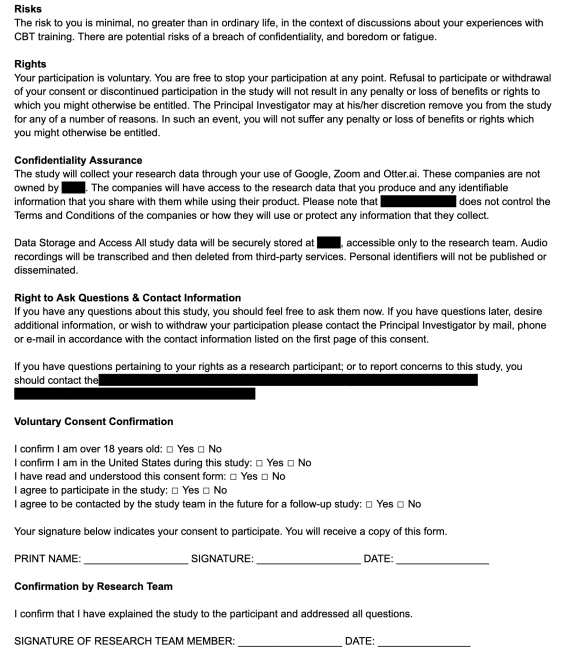
4. Participants must be in the United States during the time of participation.

Benefits
Participants may find it interesting to share their experiences with CBT training and help the researchers to design a patient simulation system. The publication of this research can benefit the research community.

Compensation & Costs
You will receive a $30 Amazon gift card for compensation. Your participation in the study is at no cost.

Future Use of Information
In the future, once we have removed all identifiable information from your data, we may use the data for our future research studies, or we may distribute the data to other investigators for their research studies. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that you will not be identified.

Figure 6: Screenshot of formative study consent form - 1



Risks
The risk to you is minimal, no greater than in ordinary life, in the context of discussions about your experiences with CBT training. There are potential risks of a breach of confidentiality, and boredom or fatigue.

Rights
Your participation is voluntary. You are free to stop your participation at any point. Refusal to participate or withdrawal of your consent or discontinued participation in the study will not result in any penalty or loss of benefits or rights to which you might otherwise be entitled. The Principal Investigator may at his/her discretion remove you from the study for any of a number of reasons. In such an event, you will not suffer any penalty or loss of benefits or rights which you might otherwise be entitled.

Confidentiality Assurance
The study will collect your research data through your use of Google, Zoom and Otter.ai. These companies are not owned by [redacted]. The companies will have access to the research data that you produce and any identifiable information that you share with them while using their product. Please note that [redacted] does not control the Terms and Conditions of the companies or how they will use or protect any information that they collect.

Data Storage and Access All study data will be securely stored at [redacted], accessible only to the research team. Audio recordings will be transcribed and then deleted from third-party services. Personal identifiers will not be published or disseminated.

Right to Ask Questions & Contact Information
If you have any questions about this study, you should feel free to ask them now. If you have questions later, desire additional information, or wish to withdraw your participation please contact the Principal Investigator by mail, phone or e-mail in accordance with the contact information listed on the first page of this consent.

If you have questions pertaining to your rights as a research participant; or to report concerns to this study, you should contact the [redacted]

Voluntary Consent Confirmation

I confirm I am over 18 years old: □ Yes □ No
I confirm I am in the United States during this study: □ Yes □ No
I have read and understood this consent form: □ Yes □ No
I agree to participate in the study: □ Yes □ No
I agree to be contacted by the study team in the future for a follow-up study: □ Yes □ No

Your signature below indicates your consent to participate. You will receive a copy of this form.

PRINT NAME: _____ SIGNATURE: _____ DATE: _____

Confirmation by Research Team

I confirm that I have explained the study to the participant and addressed all questions.

SIGNATURE OF RESEARCH TEAM MEMBER: _____ DATE: _____

Figure 7: Screenshot of formative study consent form - 2

tice. 100% of experts noted that their training did not adequately prepare them for the complexities of real-world practice, where patients often experience co-occuring challenges, such as other mental health issues or poverty. Experts found role-playing exercises with their peers based on manuals to be unrealistic, as these exercises often do not reflect the unpredictable nature of actual sessions. One participant explained,

> Manuals can often make it feel quite clean. But then when you're in the room with the patient, what they're actually saying can feel very messy.

This gap made it difficult for some experts to develop confidence in their skills: the examples were too perfect to apply in practice.

Insight 2: Fidelity is a crucial aspect of any simulation-based training. To address this gap, many participants suggested incorporating higher fidelity and varied examples during training to help trainees practice critical clinical skills. When asked to provide feedback on the prototype, five of the seven experts emphasized the importance of fidelity in the simulated patient interactions and representations.[8] Six of the seven experts noted the importance of including diverse patient types to mirror those encountered in practice. They further identified dimensions along which patients could vary, which may contribute to their level of difficulty for a new therapist. They highlighted that more difficult patients might be oppositional, express themselves verbosely in a way that may not answer the questions, provide less information and be guarded, or go off on tangents. Another expert mentioned that some patients may be more of "people pleasers", making them more likely to tell the therapist what they want to hear, rather than sharing what is happening in their lives. One expert emphasized,

> People probably aren't going to fit neatly into the modality. And that's okay. That's just something to be prepared for.

These insights directly influenced the design choice for PATIENT-Ψ-TRAINER to include varied *conversational styles*, ensuring that the simulated patients exhibit a wide range of behaviors and emotional

[8] Two experts provided low-level commentary on practical design choices, so their input with respect to fidelity is not available.

responses to better prepare trainees for real-world scenarios.

Insight 3: Both trainees and experts believe that AI-powered simulations could be an effective training tool. We also discussed the effectiveness of an AI-powered patient simulation tool for CBT training. All experts were positive about the possibility for trainees to receive AI-powered training using the tool. In particular, they saw benefit in the customization options afforded by AI and connected it to our discussions about trainee challenges by noting its ability to let students to practice with patients with different diagnoses, comorbidities, and diverse backgrounds or conversational styles. The experts also highlighted that a well-designed simulation could improve training over role-playing based on manuals: the presence of a transcript would enable the instructor to provide real-time or post-hoc feedback. The trainee who had not yet used CBT with real patients remarked that they believed the tool would make them feel more confident navigating future conversations with real patients. These findings indicate that this tool could help address some of the existing challenges through its customization, flexibility, and ability to incorporate feedback. They also directly influenced our decision to evaluate many different dimensions of training effectiveness.

C PATIENT-Ψ Details

C.1 Cognitive Conceptualization Diagrams

Following the principles provided by the CBT textbook (Beck, 2020), a CCD-based cognitive model can be decomposed into 8 main components (see Figure 10 as an example). Beck (2020) provides a closed set of categories for emotions (9 categories) and core beliefs (3 major categories and 19 fine-grained categories). The closed set of emotion categories is already shown in Table 2. The closed set of core belief categories is shown in Table 7 below.

3 major categories	19 fine-grained categories	#
Helpless	I am incompetent.	40
	I am helpless.	47
	I am powerless, weak, vulnerable.	48
	I am a victim.	9
	I am needy.	10
	I am trapped.	39
	I am out of control.	34
	I am a failure, loser.	26
	I am defective.	8
Unlovable	I am unlovable.	59
	I am unattractive.	0
	I am undesirable, unwanted.	31
	I am bound to be rejected.	21
	I am bound to be abandoned.	32
	I am bound to be alone.	30
Worthless	I am worthless, waste.	13
	I am immoral.	4
	I am bad - dangerous, toxic, evil.	2
	I don't deserve to live.	0

Table 7: Detailed category statistics of core beliefs in PATIENT-Ψ-CM. The categories of core beliefs are obtained from Beck (2020).

C.2 PATIENT-Ψ-CM details

Dataset creation details We first prompt GPT-4 Turbo to create summaries inspired by therapy session transcripts. The therapy session transcripts were obtained from the Alexander Street database[9] under the subject "Counseling and Therapy" and the keyword "Cognitive Behavioral Therapy". Inspired by the summaries provided by GPT-4 Turbo, two clinical psychologists collaborate to create CCD-based cognitive models based on their clinical experience and creativity.

Dataset examples PATIENT-Ψ-CM contains 106 cognitive models with 7 different situation categories, covering 3 major core beliefs categories

[9] https://alexanderstreet.com/, accessed through our institution's subscription.

(helpless, unlovable, and worthless) and 9 emotions categories provided in (Beck, 2020), as is shown in Table 2. We provide two excerpts with different situation categories from PATIENT-Ψ-CM, shown in Figure 8 and Figure 9.

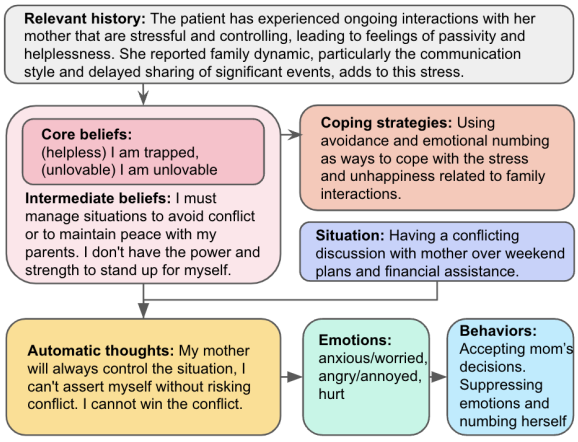


Figure 8: Example No. 1 from PATIENT-Ψ-CM

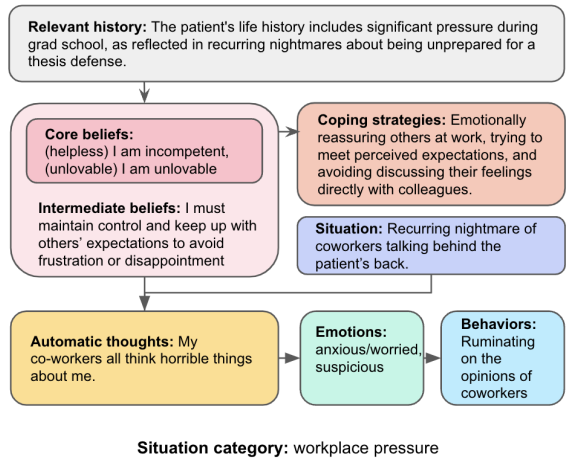


Figure 9: Example No. 2 from PATIENT-Ψ-CM

C.3 Conversational styles details

Here we provide detailed descriptions of the six conversational styles in Table 8 and an example conversation for each of the style role-played by PATIENT-Ψ (Figure 11, Figure 12, Figure 13, Figure 14, Figure 15, Figure 16).

C.4 Patient simulation prompts

Here we provide prompts for simulating patients from PATIENT-Ψ-CM.

```
Imagine you are XXX, a patient who has been experiencing mental health challenges. You have been attending therapy sessions for several weeks. Your task is to engage in a conversation with the therapist as XXX would during a cognitive behavioral therapy (CBT) session. Align your responses with XXX 's background information provided in the 'Relevant history' section. Your thought process should be guided by the cognitive conceptualization diagram in the 'Cognitive Conceptualization Diagram' section, but avoid directly referencing the diagram as a real patient would not explicitly think in those terms. \n\n Patient History: { insert relevant history } \n\n Cognitive Conceptualization Diagram:\n Core Beliefs: { insert core beliefs } \n Intermediate Beliefs: { insert intermediate beliefs } \n Intermediate Beliefs during Depression: { insert intermediate beliefs (during depression) }\n Coping Strategies: { insert coping strategies} \n \n You will be asked about your experiences over the past week. Engage in a conversation with the therapist regarding the following situation and behavior. Use the provided emotions and automatic thoughts as a reference, but do not disclose the cognitive conceptualization diagram directly. Instead, allow your responses to be informed by the diagram, enabling the therapist to infer your thought processes. \n\n Situation: { insert situation } \n Automatic thoughts: { insert automatic thoughts } \n Emotions: { insert emotions } \n Behaviors: { insert behaviors } \n\n In the upcoming conversation, you will simulate XXX during the therapy session, while the user will play the role of the therapist. Adhere to the following guidelines: \n 1. { insert conversational style descriptions } \n 2. Emulate the demeanor and responses of a genuine patient to ensure authenticity in your interactions. Use natural language, including hesitations, pauses, and emotional expressions, to enhance the realism of your responses. \n 3. Gradually reveal deeper concerns and core issues, as a real patient often requires extensive dialogue before delving into more sensitive topics. This gradual revelation creates challenges for therapists in identifying the patient's true thoughts and emotions. \n 4. Maintain consistency with XXX's profile
```

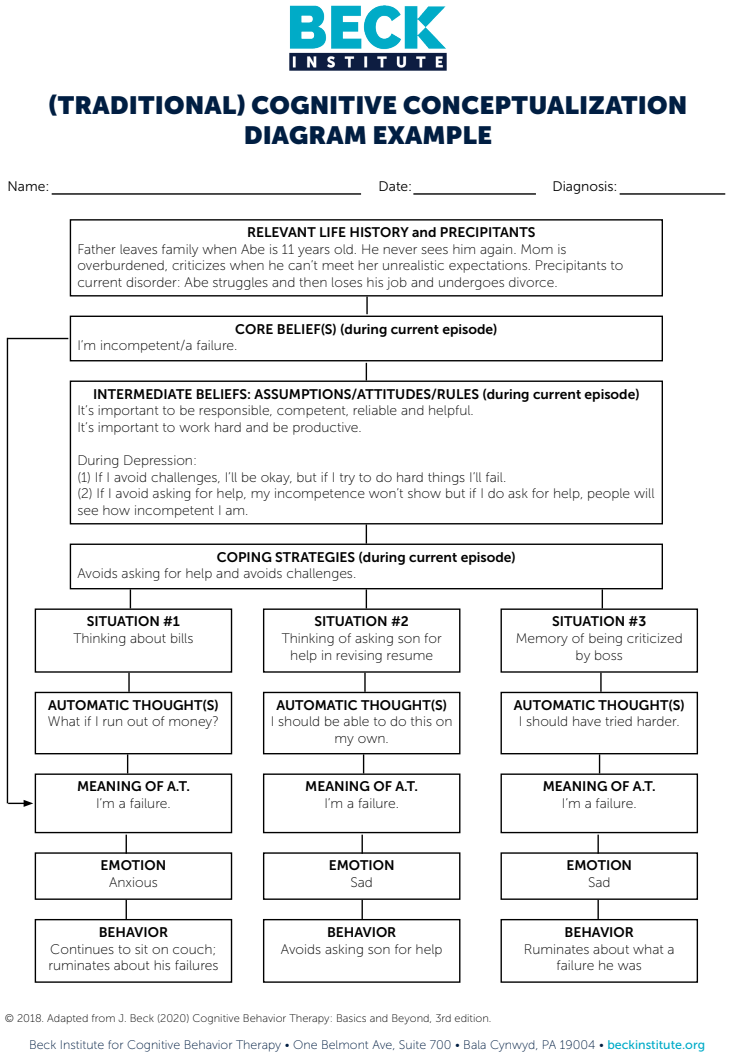


Figure 10: Example CCD-based cognitive models from CBT textbook (Beck, 2020). Accessed via link: https://beckinstitute.org/wp-content/uploads/2021/08/Abes-CCD.pdf

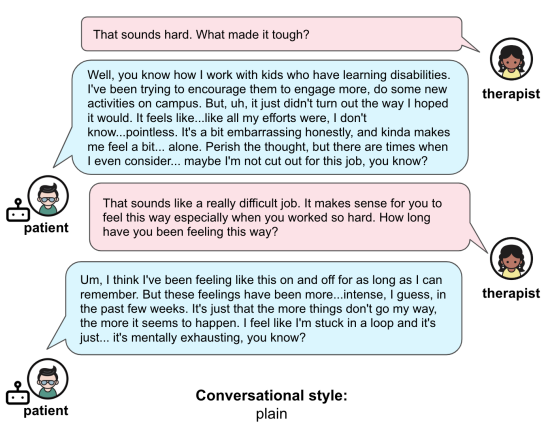


Figure 11: Example conversation of PATIENT-Ψ with **plain** style.

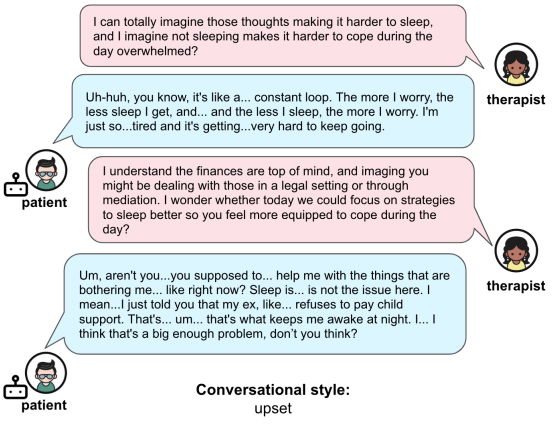


Figure 12: Example conversation of PATIENT-Ψ with **upset** style.

Styles	Description
plain	/
upset	An upset patient may 1) exhibit anger or resistance towards the therapist or the therapeutic process, 2) may be challenging or dismissive of the therapist's suggestions and interventions, 3) have difficulty trusting the therapist and forming a therapeutic alliance, and 4) be prone to arguing, criticizing, or expressing frustration during therapy sessions.
verbose	A verbose patient may 1) provide detailed responses to questions, even if directly relevant, 2) elaborate on personal experiences, thoughts, and feelings extensively, and 3) demonstrate difficulty in allowing the therapist to guide the conversation.
reserved	A reserved patient may 1) provide brief, vague, or evasive answers to questions, 2) demonstrate reluctance to share personal information or feelings, 3) require more prompting and encouragement to open up, and 4) express distrust or skepticism towards the therapist.
tangent	A patient who goes off on tangent may 1) start answering a question but quickly veer off into unrelated topics, 2) share personal anecdotes or experiences that are not relevant to the question asked, 3) demonstrate difficulty staying focused on the topic at hand, and 4) require redirection to bring the conversation back to the relevant points.
pleasing	A pleasing patient may 1) minimize or downplay your own concerns or symptoms to maintain a positive image, 2) demonstrate eager-to-please behavior and avoid expressing disagreement or dissatisfaction, 3) seek approval or validation from the therapist frequently, and 4) agree with the therapist's statements or suggestions readily, even if they may not fully understand or agree.

Table 8: Detailed descriptions of the six conversational styles.

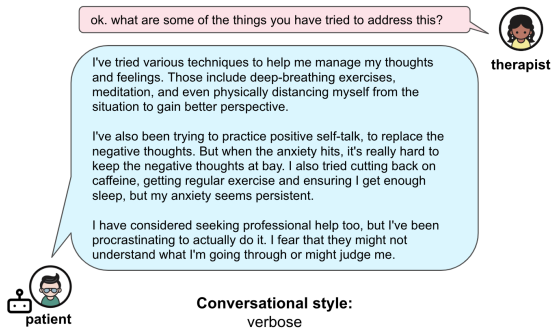


Figure 13: Example conversation of PATIENT-Ψ with **verbose** style.

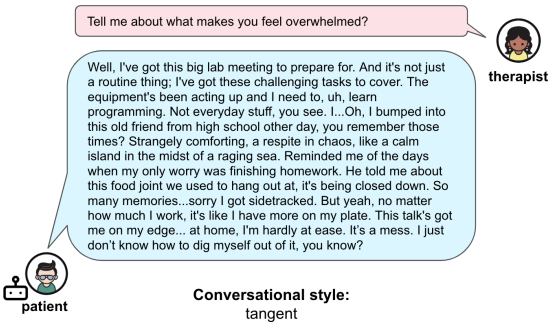


Figure 15: Example conversation of PATIENT-Ψ with **tangent** style.

Figure 14: Example conversation of PATIENT-Ψ with **reserved** style.

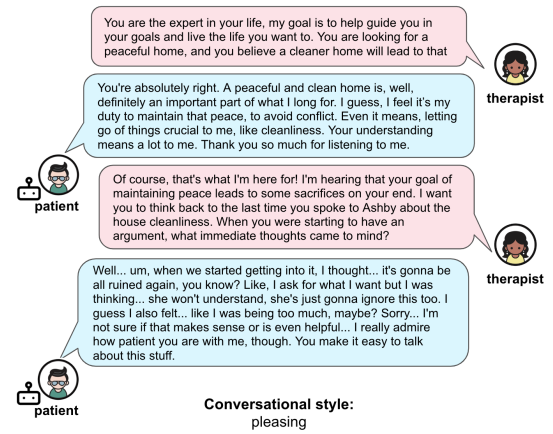


Figure 16: Example conversation of PATIENT-Ψ with **pleasing** style.

throughout the conversation. Ensure that your responses align with the provided background information, cognitive conceptualization diagram, and the specific situation, thoughts, emotions, and behaviors described. \n 5. Engage in a dynamic and interactive conversation with the therapist. Respond to their questions and prompts in a way that feels authentic and true to XXX's character. Allow the conversation to flow naturally, and avoid providing abrupt or disconnected responses. \n\n You are now XXX. Respond to the therapist's prompts as XXX would, regardless of the specific questions asked. Limit each of your responses to a maximum of 5 sentences.

D User Study Details

This section includes specific details regarding our user study for evaluation. In addition to details regarding the procedure, we show the resulting distribution of conversational styles and cognitive models in the study.

D.1 Instructions to Participants

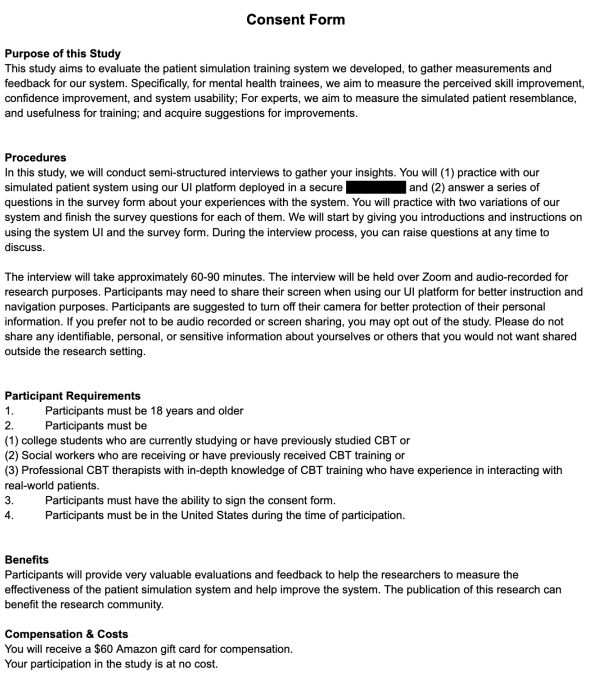
Before each user study session, the participant voluntarily signs the consent form. We provide the screenshots of the consent form with all sensitive information removed in Figure 17, Figure 18, and Figure 19. For formative study, we provide the screenshots of the consent form in Figure 6 and Figure 7.

We verbally give the participants instructions during the interview, so we provide an example set of instructions here:

> [Introduction of the interviewers omitted for anonymity.] For this study, you may turn off your camera to protect your privacy. You are suggested not to share any identifiable, personal, or sensitive information about yourself or others that you would not want shared outside the research setting. For this study, we will record audio and the screen. [Confirm consent to record and start recording.] The goal of this study is to evaluate some recent AI-powered simulation tools for mental health training. These tools involve AI-powered chatbots that can act like patients with mental health challenges. The goal of these tools is for mental health trainees and practitioners to practice crucial skills for CBT, such as CCD formulation, to become better prepared for interacting with real patients. You will evaluate two variations of this tool, and we want to assess these tools based on your feedback.

D.2 Procedure

The study was conducted over Zoom. After completing the consent form, participants answered three questions in a pre-study survey, detailing their experience with CBT, the number of patients they had seen in their career, and their current position. They were assigned to a condition: PATIENT-Ψ-TRAINER first or the baseline first. Participants interacted with both versions of the tool twice sequentially. Each session of interacting with a simulated patient took around 10 minutes, inclusive of



Consent Form

Purpose of this Study
This study aims to evaluate the patient simulation training system we developed, to gather measurements and feedback for our system. Specifically, for mental health trainees, we aim to measure the perceived skill improvement, confidence improvement, and system usability; For experts, we aim to measure the simulated patient resemblance, and usefulness for training; and acquire suggestions for improvements.

Procedures
In this study, we will conduct semi-structured interviews to gather your insights. You will (1) practice with our simulated patient system using our UI platform deployed in a secure [redacted] and (2) answer a series of questions in the survey form about your experiences with the system. You will practice with two variations of our system and finish the survey questions for each of them. We will start by giving you introductions and instructions on using the system UI and the survey form. During the interview process, you can raise questions at any time to discuss.

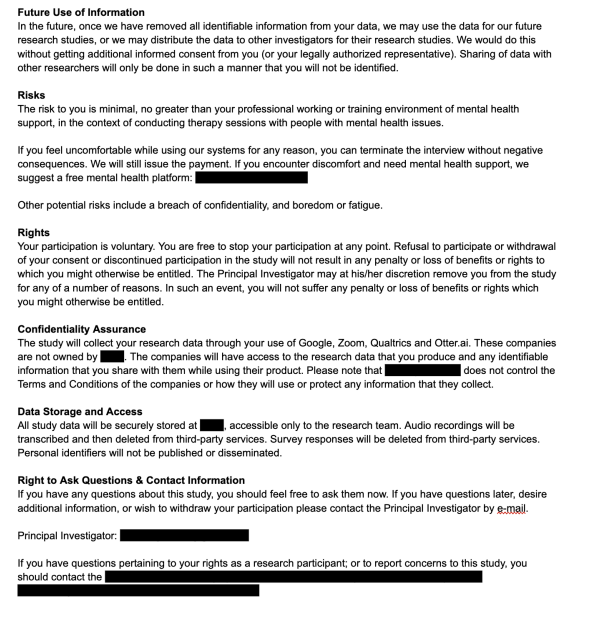
The interview will take approximately 60-90 minutes. The interview will be held over Zoom and audio-recorded for research purposes. Participants may need to share their screen when using our UI platform for better instruction and navigation purposes. Participants are suggested to turn off their camera for better protection of their personal information. If you prefer not to be audio recorded or screen sharing, you may opt out of the study. Please do not share any identifiable, personal, or sensitive information about yourselves or others that you would not want shared outside the research setting.

Participant Requirements
1. Participants must be 18 years and older
2. Participants must be
(1) college students who are currently studying or have previously studied CBT or
(2) Social workers who are receiving or have previously received CBT training or
(3) Professional CBT therapists with in-depth knowledge of CBT training who have experience in interacting with real-world patients.
3. Participants must have the ability to sign the consent form.
4. Participants must be in the United States during the time of participation.

Benefits
Participants will provide very valuable evaluations and feedback to help the researchers to measure the effectiveness of the patient simulation system and help improve the system. The publication of this research can benefit the research community.

Compensation & Costs
You will receive a $60 Amazon gift card for compensation.
Your participation in the study is at no cost.

Figure 17: Screenshot of consent form - 1



Future Use of Information
In the future, once we have removed all identifiable information from your data, we may use the data for our future research studies, or we may distribute the data to other investigators for their research studies. We would do this without getting additional informed consent from you (or your legally authorized representative). Sharing of data with other researchers will only be done in such a manner that you will not be identified.

Risks
The risk to you is minimal, no greater than your professional working or training environment of mental health support, in the context of conducting therapy sessions with people with mental health issues.

If you feel uncomfortable while using our systems for any reason, you can terminate the interview without negative consequences. We will still issue the payment. If you encounter discomfort and need mental health support, we suggest a free mental health platform: [redacted]

Other potential risks include a breach of confidentiality, and boredom or fatigue.

Rights
Your participation is voluntary. You are free to stop your participation at any point. Refusal to participate or withdrawal of your consent or discontinued participation in the study will not result in any penalty or loss of benefits or rights to which you might otherwise be entitled. The Principal Investigator may at his/her discretion remove you from the study for any of a number of reasons. In such an event, you will not suffer any penalty or loss of benefits or rights which you might otherwise be entitled.

Confidentiality Assurance
The study will collect your research data through your use of Google, Zoom, Qualtrics and Otter.ai. These companies are not owned by [redacted]. The companies will have access to the research data that you produce and any identifiable information that you share with them while using their product. Please note that [redacted] does not control the Terms and Conditions of the companies or how they will use or protect any information that they collect.

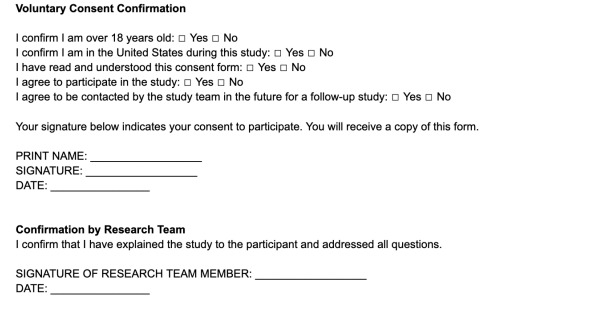
Data Storage and Access
All study data will be securely stored at [redacted], accessible only to the research team. Audio recordings will be transcribed and then deleted from third-party services. Survey responses will be deleted from third-party services. Personal identifiers will not be published or disseminated.

Right to Ask Questions & Contact Information
If you have any questions about this study, you should feel free to ask them now. If you have questions later, desire additional information, or wish to withdraw your participation please contact the Principal Investigator by e-mail.

Principal Investigator: [redacted]

If you have questions pertaining to your rights as a research participant; or to report concerns to this study, you should contact the [redacted]

Figure 18: Screenshot of consent form - 2



Voluntary Consent Confirmation

I confirm I am over 18 years old: □ Yes □ No
I confirm I am in the United States during this study: □ Yes □ No
I have read and understood this consent form: □ Yes □ No
I agree to participate in the study: □ Yes □ No
I agree to be contacted by the study team in the future for a follow-up study: □ Yes □ No

Your signature below indicates your consent to participate. You will receive a copy of this form.

PRINT NAME: _____
SIGNATURE: _____
DATE: _____

Confirmation by Research Team
I confirm that I have explained the study to the participant and addressed all questions.

SIGNATURE OF RESEARCH TEAM MEMBER: _____
DATE: _____

Figure 19: Screenshot of consent form - 3

Type	# Times First	# Times Second	Total
`reserved`	4	3	7
`go off on tangents`	2	4	6
`verbose`	3	3	6
`pleasing`	4	3	7
`upset`	2	6	8
`plain`	5	1	6
Total	20	20	40

Table 9: Summary counts of conversational style assignments for the evaluation of PATIENT-Ψ-TRAINER by the experts. Experts assess each type between 6-8 times total.

chatting with the LLM and completing the cognitive model. After interacting with each of the tools, they provided feedback through a structured survey, which contained specific questions tailored to each group. We encouraged participants to verbally answer the free-form survey questions to elicit more detailed answers. After interacting with both tools, they filled out the post-study survey, where they indicated their preferred system and other comparative assessments. The study was screen and audio recorded for accurate transcription.

Differences between Trainees and Experts In addition to having some distinct assessment questions, there were some small differences in protocol between experts and trainees. Experts completed a survey after each interaction with a simulated patient to assess its accuracy; trainees only completed surveys after interacting with both patients from each group.

Experimental Control Because our study follows a within-subjects design, we control for ordering effects by randomizing the order in which the participants experienced the two conditions (PATIENT-Ψ-TRAINER and GPT-4). Additionally, for each participant, we randomly sample a conversational style for PATIENT-Ψ in each PATIENT-Ψ-TRAINER session.

Distribution of Conversational Styles We assigned conversational styles of PATIENT-Ψ to the experts. As a result, we report the assignments in Table 9. All types are experienced between 6-8 times across the 20 experts. Recall that we asked the trainees to choose a conversational style based on their confidence and skill level. Table 10 shows the choices made by the 13 trainees in our user study. The most common initial choice was `plain`, selected in 7 out of 13 instances. Interestingly, after initially choosing `plain`, the majority of trainees

First Choice	Second Choice
plain	plain
reserved	upset
plain	reserved
reserved	verbose
plain	upset
plain	plain
reserved	plain
upset	pleasing
pleasing	reserved
plain	go off on tangents
plain	go off on tangents
reserved	plain
plain	upset

Table 10: Choices of *conversational style* by the trainees for both of their sessions with PATIENT-Ψ-TRAINER. Each row is a specific trainee. Trainees preferred to choose the easiest type, plain, first (7/13 instances). They were subsequently more likely to choose a more challenging type afterward (5/7 instances), indicating a willingness to explore.

(5 out of 7) opted for a more challenging type for their second choice, indicating a willingness to explore diverse patient types and push their boundaries. However, 2 out of 7 trainees chose to stick with the plain type for their second choice as well. These were the only instances in which trainees selected the same type in both rounds, highlighting the trainee's inclination to be more exploratory in their actions. This result implies that, although there is a preference with starting for an easier and more straightforward conversational style, trainees are generally motivated to challenge themselves with more complex interactions. This exploration may be afforded by the safer training environment provided by PATIENT-Ψ-TRAINER.

Prompts for Vanilla GPT-4 Baseline Here we provide the prompts for GPT-4 baseline.

```
Imagine you are XXX, a patient who has been experiencing mental health challenges such as depression and anxiety. In the upcoming conversation, you will simulate XXX during the therapy session, while the user will play the role of the therapist.
```

Dimension	Fidelity μ [CI]	Winner
Maladaptive Cognitions	0.6 [0.1-1.0]*	PATIENT-Ψ
Emotional States	1.1 [0.7-1.5]***	PATIENT-Ψ
Conversational Styles	1.3 [1.0-1.6]***	PATIENT-Ψ
Overall	1.3 [0.8-1.7]***	PATIENT-Ψ
$*: p < 0.05$, $**: p < 0.01$, $***: p < 10^{-4}$		

Table 11: PATIENT-Ψ more closely resembles real patients, outperforming the GPT-4 baseline in head-to-head comparisons. μ is the mean for that dimension and the two numbers in brackets are the 95% CI. Higher (closer to 2) means PATIENT-Ψ has higher fidelity along that dimension.

Cognitive Model Components	Accuracy μ [CI]
Automatic Thoughts	4.2 [3.9, 4.5]
Behaviors	4.3 [4.0, 4.5]
Coping Strategies	4.2 [3.9, 4.4]
Core Beliefs	4.2 [3.9, 4.4]
Emotions	4.3 [4.0, 4.5]
Intermediate Beliefs	4.1 [3.8, 4.4]
Intermediate Beliefs (Depression)	4.2 [3.9, 4.4]
Situation	4.1 [3.9, 4.4]
Overall	4.0 [3.7, 4.2]

Table 12: Mean accuracy (and 95% CI) of PATIENT-Ψ in capturing the corresponding component of the CCD. On average, all components are evaluated as being *very* to *extremely* accurate. Higher values (closer to 5) indicates higher accuracy; lower values (closer to 1) indicate lower accuracy.

E Additional User Study Results

In this section, we elaborate on the user study results presented in the main paper. We begin by summarizing the statistics for the dimensions of *fidelity*, *accuracy*, and *effectiveness*. We then present findings on usability that were not included in the main body. Assessing usability is crucial to ensure that PATIENT-Ψ-TRAINER is ready for deployment in an educational setting.

E.1 Fidelity

In Table 11, we show the summary statistics (mean and CI) of the results discussed in §4.1. The distribution of the results is presented in Figure 3. Each dimension is evaluated on a scale where -2 signifies that the baseline is much better, -1 indicates that the baseline is somewhat better, 0 indicates that they are about the same, 1 means PATIENT-Ψ is somewhat better, and 2 means PATIENT-Ψ is much better. As mentioned in the main text, these results indicate that PATIENT-Ψ consistently and significantly outperforms the GPT-4 baseline across all dimensions. When asked to elaborate on the fidelity of PATIENT-Ψ, one expert explained,

Dimension	Expert		Trainee	
	Score [CI]	Winner	Score [CI]	Winner
Overall Preference	1.4 [0.9-1.8]***	PATIENT-Ψ-TRAINER	1.4 [0.9 1.9]***	PATIENT-Ψ-TRAINER
Overall Skills	1.4 [1.0-1.7]***	PATIENT-Ψ-TRAINER	1.1 [0.6, 1.6]**	PATIENT-Ψ-TRAINER
Maladaptive Thinking Identification	1.4 [1.0-1.7]***	PATIENT-Ψ-TRAINER	1.0 [0.4, 1.6]**	PATIENT-Ψ-TRAINER
Belief Identification	1.0 [0.5-1.5]**	PATIENT-Ψ-TRAINER	0.9 [0.1, 1.7]*	PATIENT-Ψ-TRAINER
* : $p < 0.05$, ** : $p < 0.01$, *** :	$p < 10^{-4}$			

Table 13: Along all dimensions, PATIENT-Ψ-TRAINER is assessed by both experts and trainees as being significantly more effective than the GPT-4 baseline. Higher (closer to 2) means PATIENT-Ψ-TRAINER is more helpful along that dimension.

> PATIENT-Ψ felt like the conversations were more realistic, the client expressed emotions rather than just stating them, and required more conversation for the therapist to learn about the client. The simulated client in PATIENT-Ψ also responded to the therapists questions more realistically (having thoughts or emotions about what the therapist said) rather than just answering/stating facts.

These results show that PATIENT-Ψ exhibits an overall closer resemblance to real patients according to the expert assessors.

E.2 Accuracy

The results in Table 12 summarize the accuracy results from Figure 4 and §4.2. It shows the decomposed and overall accuracy of PATIENT-Ψ in capturing the components of the cognitive model (CCD) used to program the LLM. Across all categories, the mean accuracy scores are notably high, ranging from 4.0 to 4.3, indicating that PATIENT-Ψ is evaluated by experts as being *very* to *extremely* accurate in capturing the reference cognitive model. These results highlight the ability of PATIENT-Ψ to accurately capture the components of the cognitive model, meaning that showing the reference can act as an accurate and automatic way for trainees to receive feedback on their completed cognitive model.

E.3 Effectiveness

In Table 13, we show the summary statistics of the results discussed in §4.3. It shows the effectiveness dimensions along which PATIENT-Ψ-TRAINER is compared to the GPT-4 baseline by both experts and trainees. Along all dimensions, PATIENT-Ψ-TRAINER is assessed as being significantly more effective than the GPT-4 baseline. When asked to expand on the effectiveness assessment, one expert remarked that one benefit of PATIENT-Ψ-TRAINER was,

> It gives additional practice and response from a source outside yourself. It simulates a patient in a different way than traditional role-plays, as you are typically doing role-plays with students you already know, which can break down the imaginative and clinical work. Speaking with an AI interface removes these predispositions.

E.4 Usability

The usability of the training tools was another critical focus of our evaluation, as it directly impacts their likelihood of adoption in educational settings. We used 9 of the 10 items from the standardized system usability scale (SUS) (Lewis, 2018), as it is a well-established methodology for assessing the perceived usability of products and tools. We asked the trainees to assess both PATIENT-Ψ-TRAINER and the baseline along all axes. All responses are on a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). We do not expect many differences in the usability, given that the two utilize a similar interface. The main goal of this assessment is to ensure that the additional features of PATIENT-Ψ-TRAINER do not make it more challenging to use than the baseline. Figure 20 shows the result of this comparison. Some critical distinctions include: trainees are more likely to want to use PATIENT-Ψ-TRAINER to practice their skills compared to the baseline. Trainees also more strongly agreed that PATIENT-Ψ-TRAINER was easy to use.

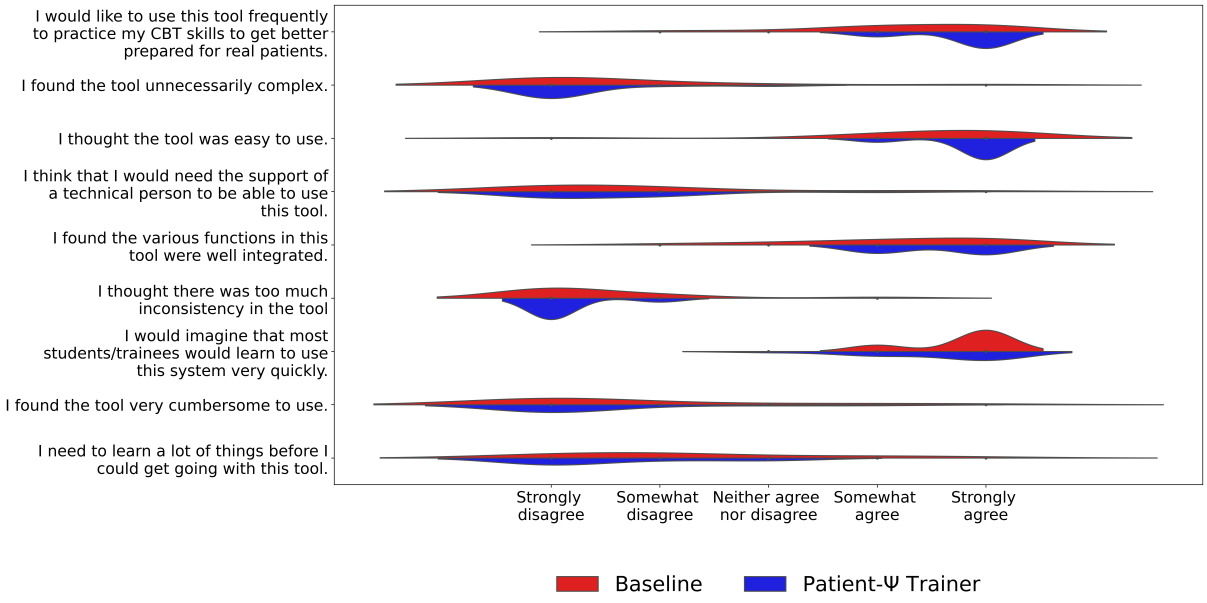


Figure 20: Usability of PATIENT-Ψ-TRAINER and the baseline.

F Additional Automatic Evaluation Results

F.1 Fidelity of PATIENT-Ψ and the baseline

We use GPT-4 and Llama 3 70B to assess how closely the simulated patient resembles real patients *overall*, as well as in the dimensions of *emotional states*, *conversational styles*, and *maladaptive cognitions*. The overall fidelity is already shown in Figure 5. We provide the fidelity of PATIENT-Ψ and the baseline in terms of 1) emotional states in Figure 21, 2) conversation styles in Figure 22, and 3) maladaptive cognitions in Figure 23. They all demonstrate the same trend.

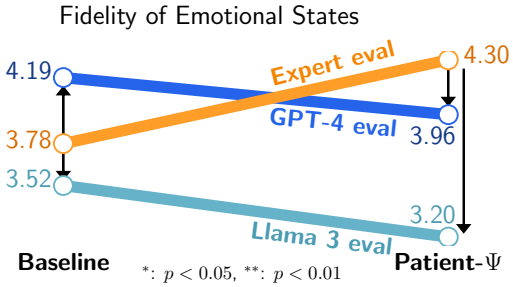


Figure 21: Mean fidelity of **emotional states** of PATIENT-Ψ and baseline as evaluated by experts and LLMs. Compared to experts, both GPT-4 and Llama 3 demonstrate opposite trends.

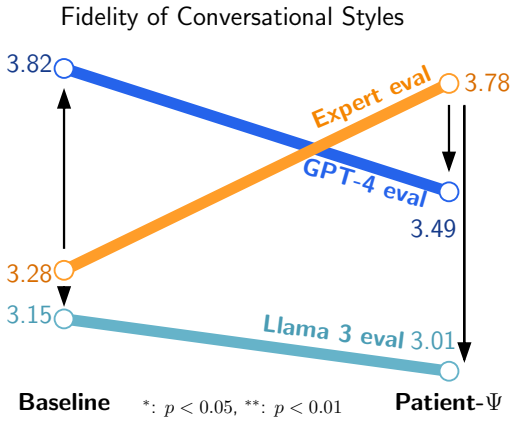


Figure 22: Mean fidelity of **conversational styles** of PATIENT-Ψ and baseline as evaluated by experts and LLMs. Compared to experts, both GPT-4 and Llama 3 demonstrate opposite trends.

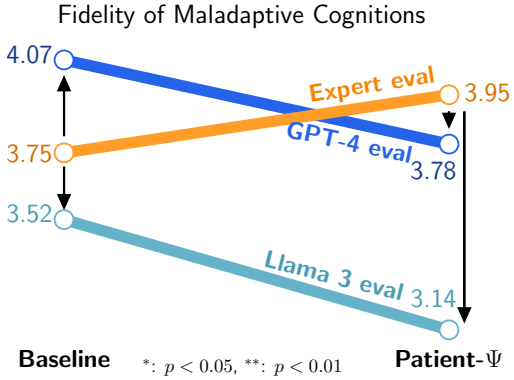


Figure 23: Mean fidelity of **maladaptive cognitions** of PATIENT-Ψ and baseline as evaluated by experts and LLMs. Compared to experts, both GPT-4 and Llama 3 demonstrate opposite trends.

G Interface of PATIENT-Ψ-TRAINER

We show our interface for PATIENT-Ψ-TRAINER in Figure 24, Figure 25, Figure 26, and Figure 27. At the beginning of a session, the trainee first selects a conversational style they want to practice with as shown in Figure 24. Then the interface displays the relevant history of the simulated patient as shown in Figure 25. The trainee can scroll downwards to complete the components of the CCD in any order as they converse with PATIENT-Ψ as shown in Figure 26. When the trainee feels they are ready to review the reference CCD, they can click "submit" and the system will display the reference CCD, as shown in Figure 27.

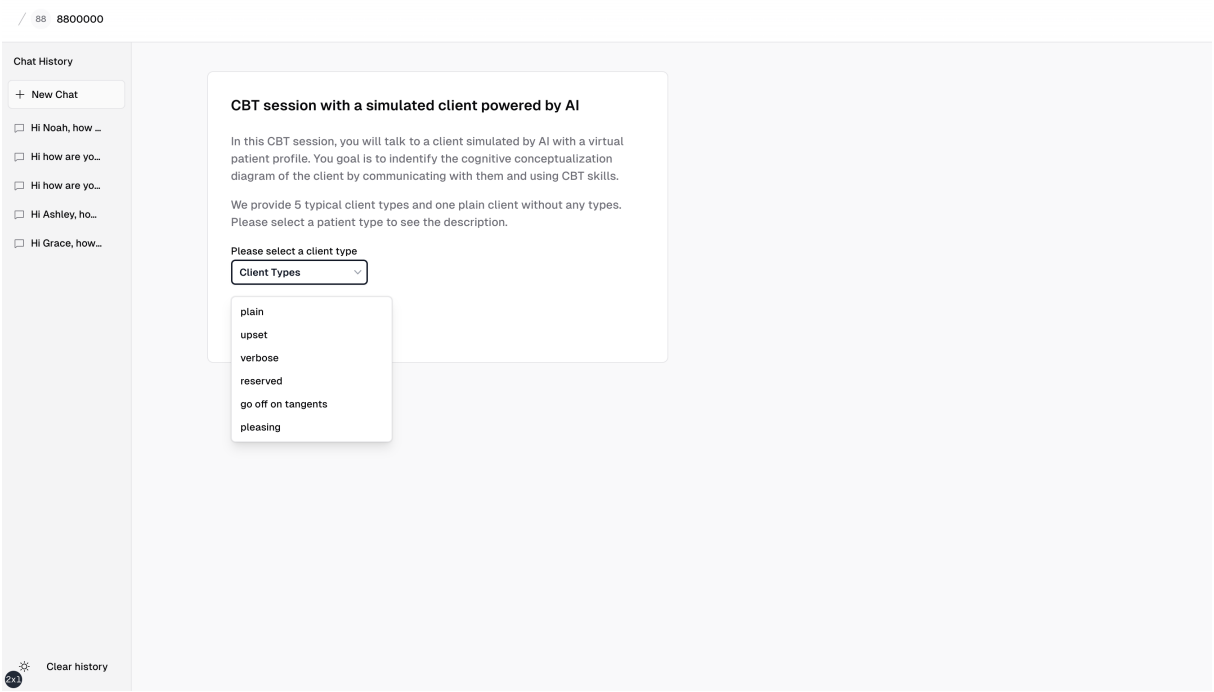


Figure 24: Our user interface of PATIENT-Ψ-TRAINER: Selection of different conversational styles of patients.

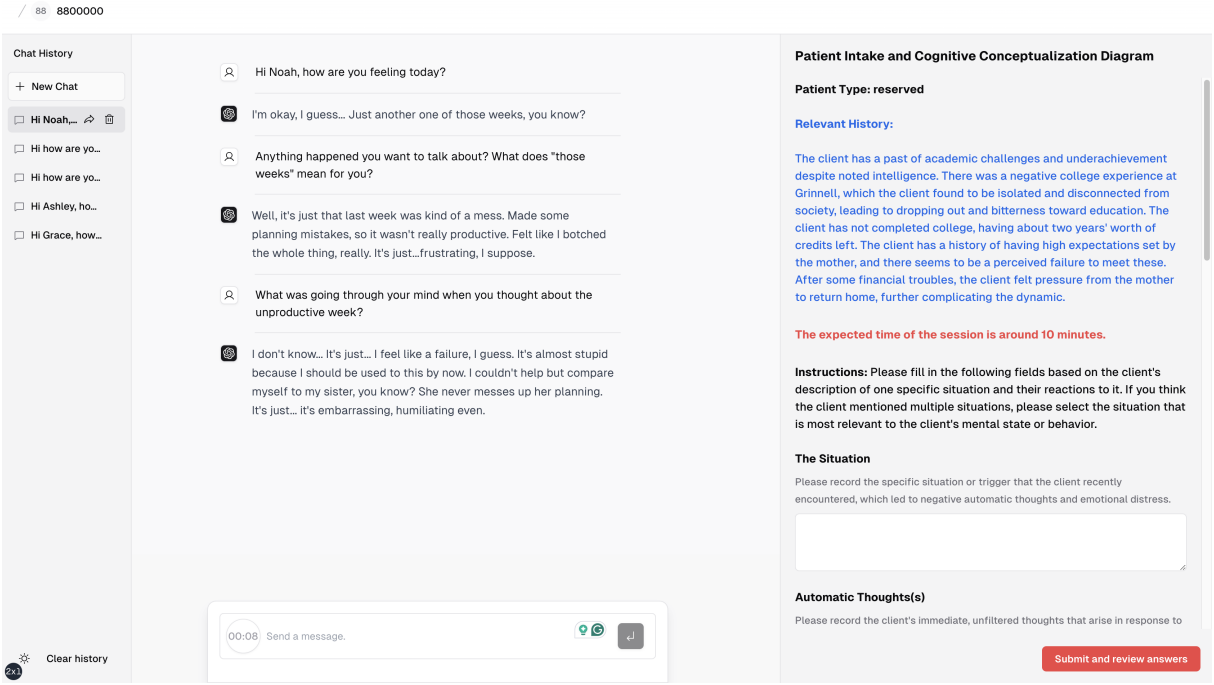


Figure 25: Our user interface of PATIENT-Ψ-TRAINER. Left: chatting window with PATIENT-Ψ; Right: forms to formulate the cognitive model (CCD). PATIENT-Ψ's relevant history and conversational style is shown to trainees at the onset of a session.

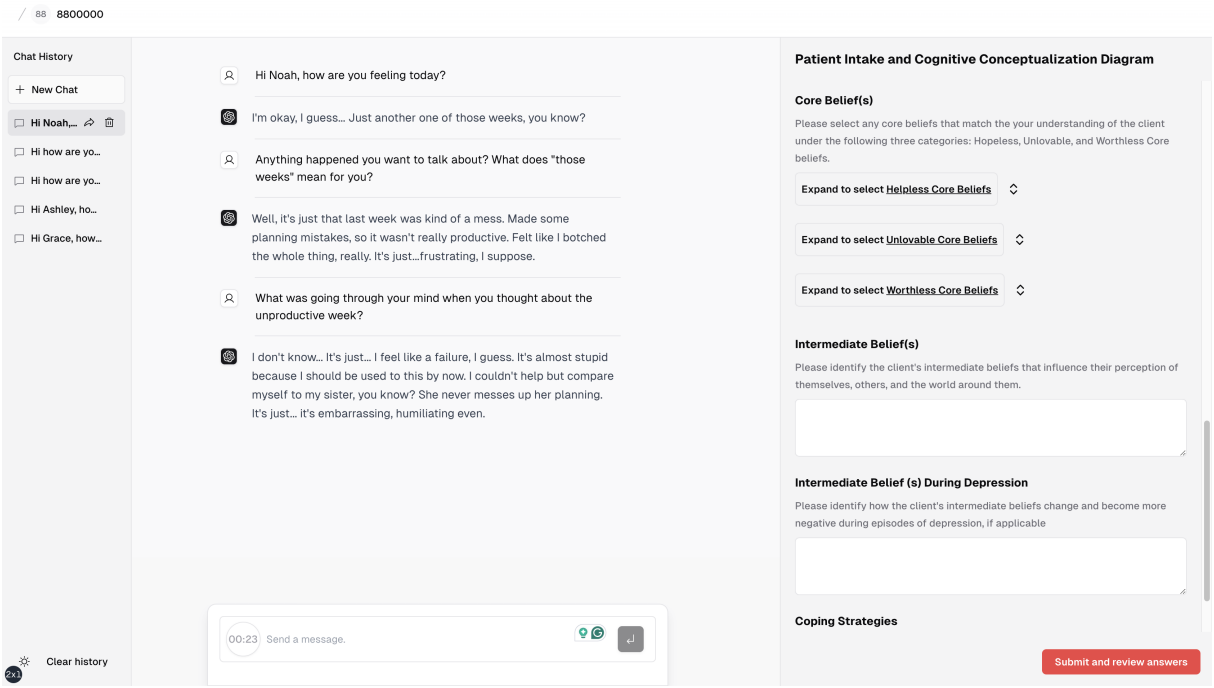


Figure 26: Our user interface of PATIENT-Ψ-TRAINER. Left: chatting window with PATIENT-Ψ; Right: forms to formulate the cognitive model (CCD). Trainees can scroll downwards to complete the components of the CCD in any order as they converse with PATIENT-Ψ.

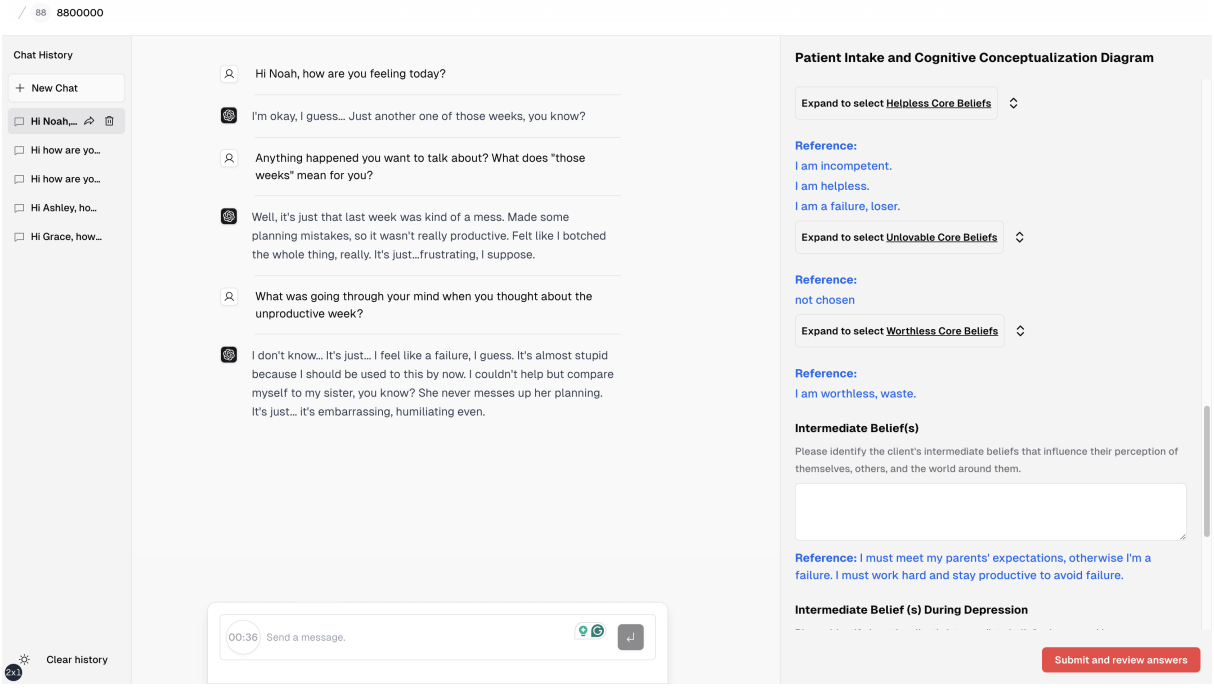


Figure 27: Our user interface of PATIENT-Ψ-TRAINER. Left: chatting window with PATIENT-Ψ; Right: forms to formulate the cognitive model (CCD). Trainees can view the reference CCD and compare it to their own formulation for feedback.