A Ethics Statement

IRB (Institutional Review Board) Approval. This project is approved by our Institutional Review Board (IRB). 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. For both the formative study and user study, we recruited participants through professional networks and 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[5], 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 Identifiable 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 after the publication of our paper.

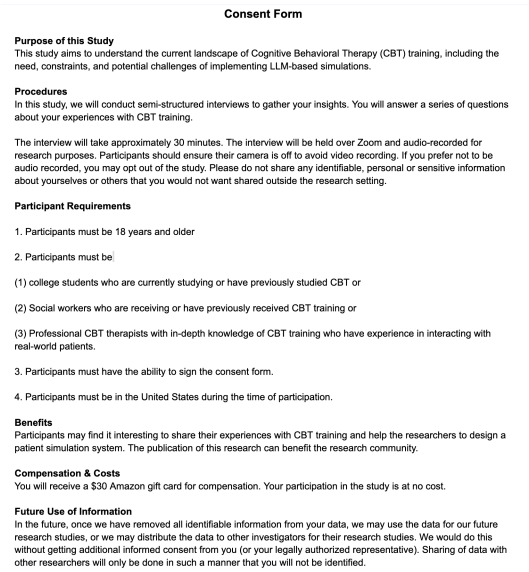
[5] `https://alexanderstreet.com/`

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.[6] 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.



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.

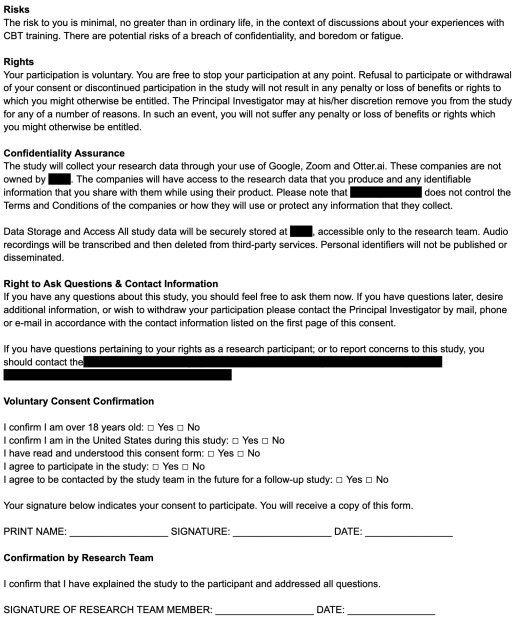
Figure 6: Screenshot of formative study consent form - 1

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 practice. 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

[6] We recruited participants through professional networks.



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

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.[7] 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 responses to better prepare trainees for real-world scenarios.

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

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[8] 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 (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.

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).

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

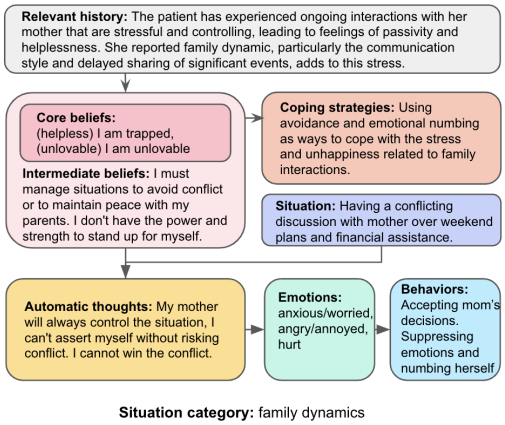


Figure 8: Example No. 1 from PATIENT-Ψ-CM

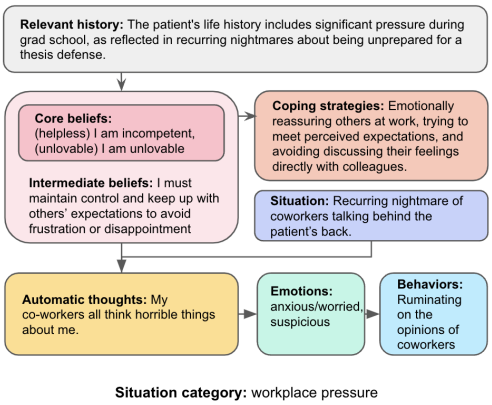


Figure 9: Example No. 2 from PATIENT-Ψ-CM

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.

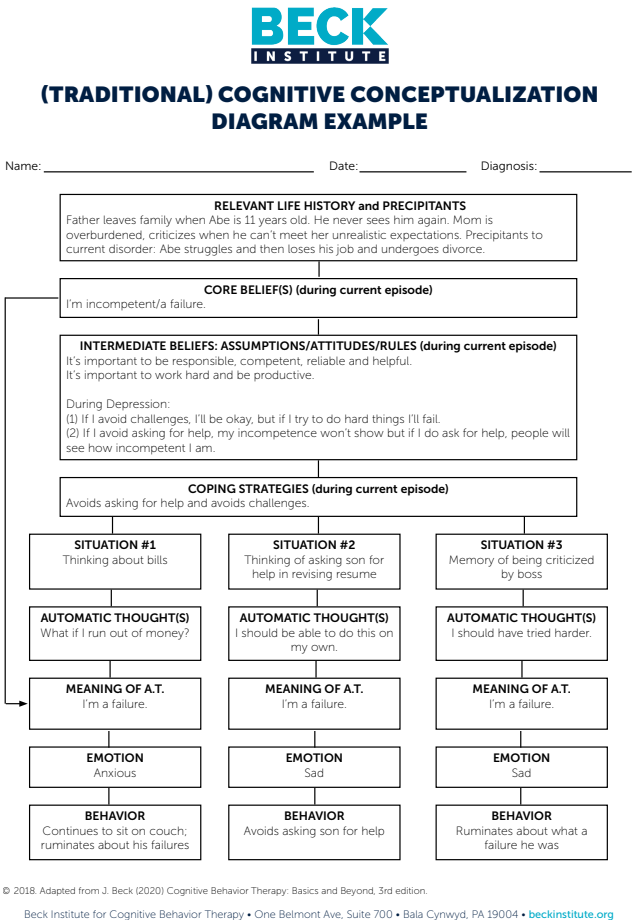


Figure 10: Example CCD-based cognitive models from CBT textbook [Beck, 2020]. Accessed via link: https://beckinstitute.org/wp-content/uploads/2021/08/Abes-CCD.pdf

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
```

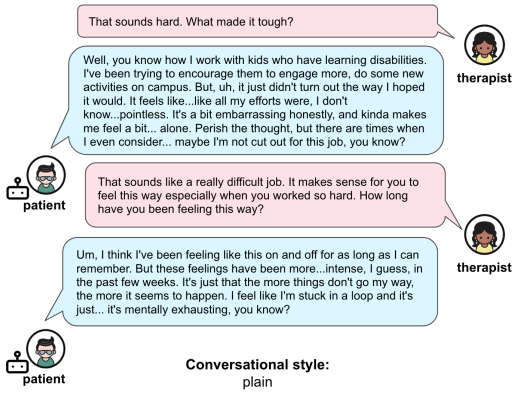


Figure 11: Example conversation of PATIENT-Ψ with **plain** style.

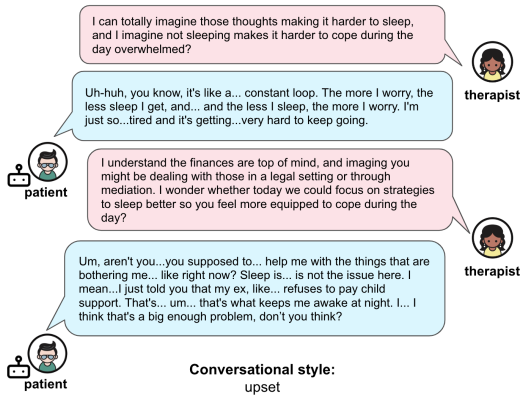


Figure 12: Example conversation of PATIENT-Ψ with **upset** style.

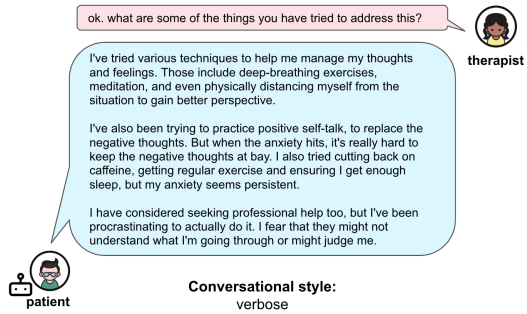


Figure 13: Example conversation of PATIENT-Ψ with **verbose** style.

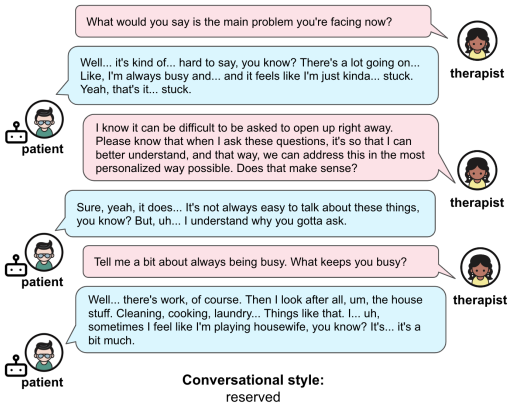


Figure 14: Example conversation of PATIENT-Ψ with **reserved** style.

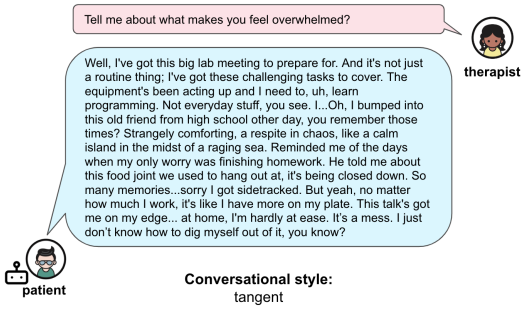


Figure 15: Example conversation of PATIENT-Ψ with **tangent** style.

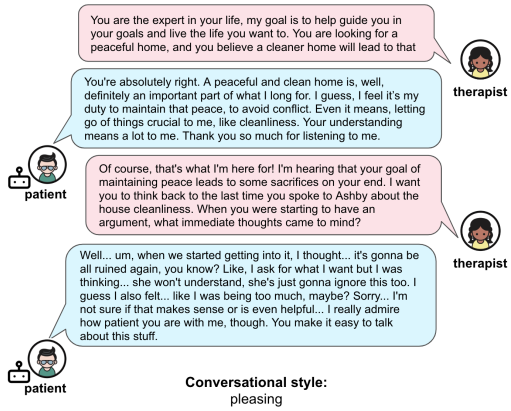


Figure 16: Example conversation of PATIENT-Ψ with **pleasing** style.

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 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

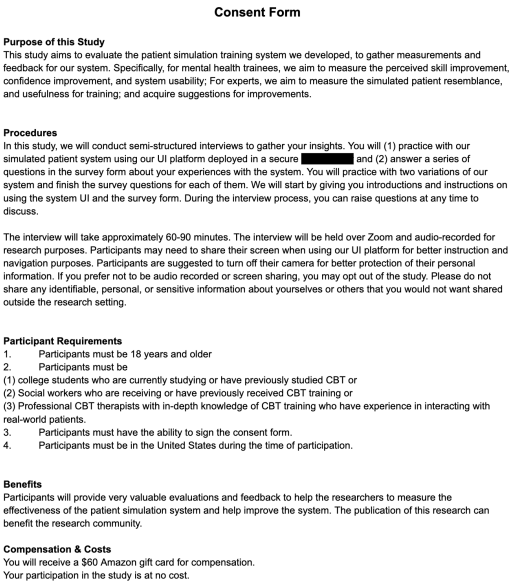
This section includes specific details regarding our user study for evaluation. In addition to details regarding the procedure, we show the resulting distribution of conversational styles and cognitive models in the study.

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.



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.

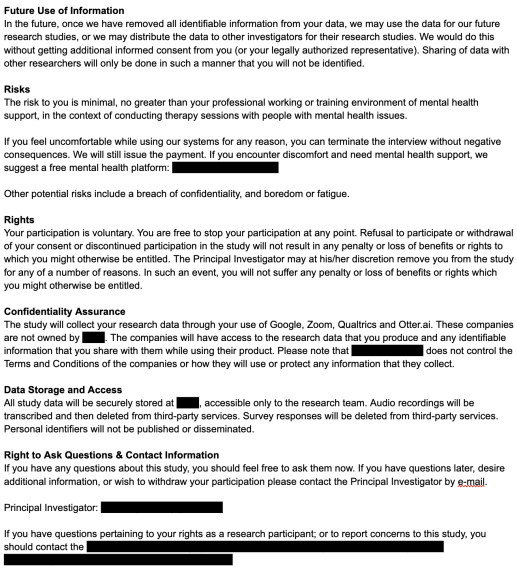
Benefits
Participants will provide very valuable evaluations and feedback to help the researchers to measure the effectiveness of the patient simulation system and help improve the system. The publication of this research can benefit the research community.

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

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



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.

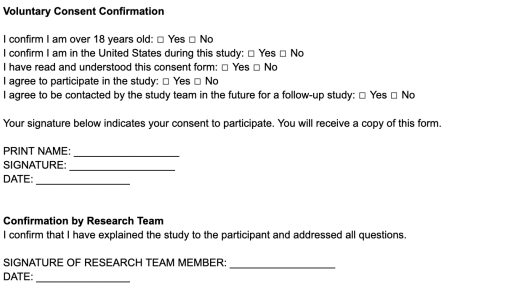
Data Storage and Access
All study data will be securely stored at [redacted], accessible only to the research team. Audio recordings will be transcribed and then deleted from third-party services. Survey responses will be deleted from third-party services. Personal identifiers will not be published or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

of 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,

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.

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.

after initially choosing plain, the majority of trainees (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

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.

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,

> 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.

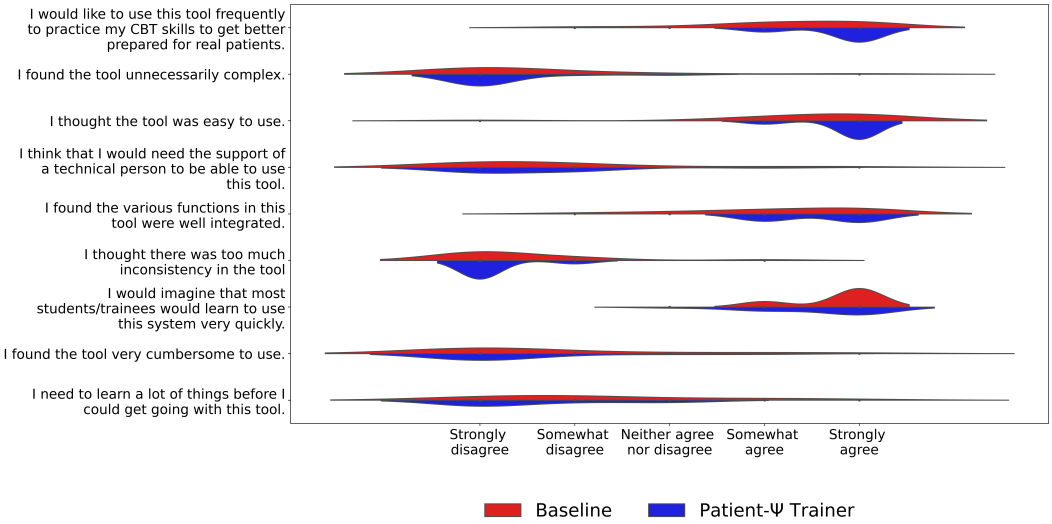


Figure 20: Usability of PATIENT-Ψ-TRAINER and the baseline.

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.

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.

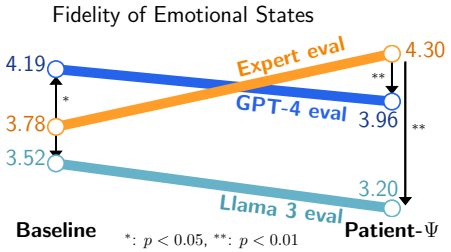


Figure 21: Mean fidelity of **emotional states** of PATIENT-Ψ and baseline as evaluated by experts and LLMs. Compared to experts, both GPT-4 and Llama 3 demonstrate opposite trends.

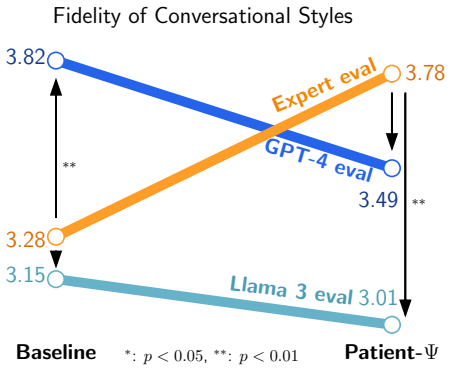


Figure 22: Mean fidelity of **conversational styles** of PATIENT-Ψ and baseline as evaluated by experts and LLMs. Compared to experts, both GPT-4 and Llama 3 demonstrate opposite trends.

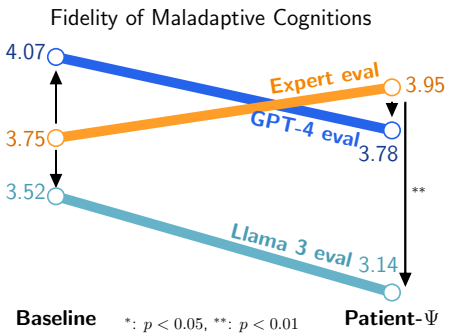


Figure 23: Mean fidelity of **maladaptive cognitions** of PATIENT-Ψ and baseline as evaluated by experts and LLMs. Compared to experts, both GPT-4 and Llama 3 demonstrate opposite trends.

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.

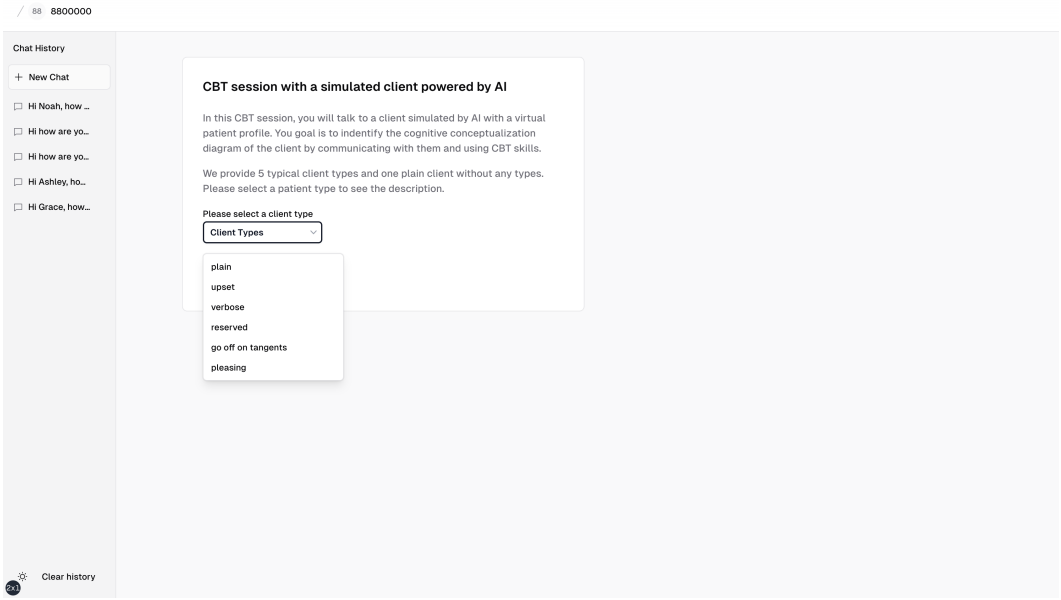


Figure 24: Our user interface of PATIENT-Ψ-TRAINER: Selection of different conversational styles of patients.

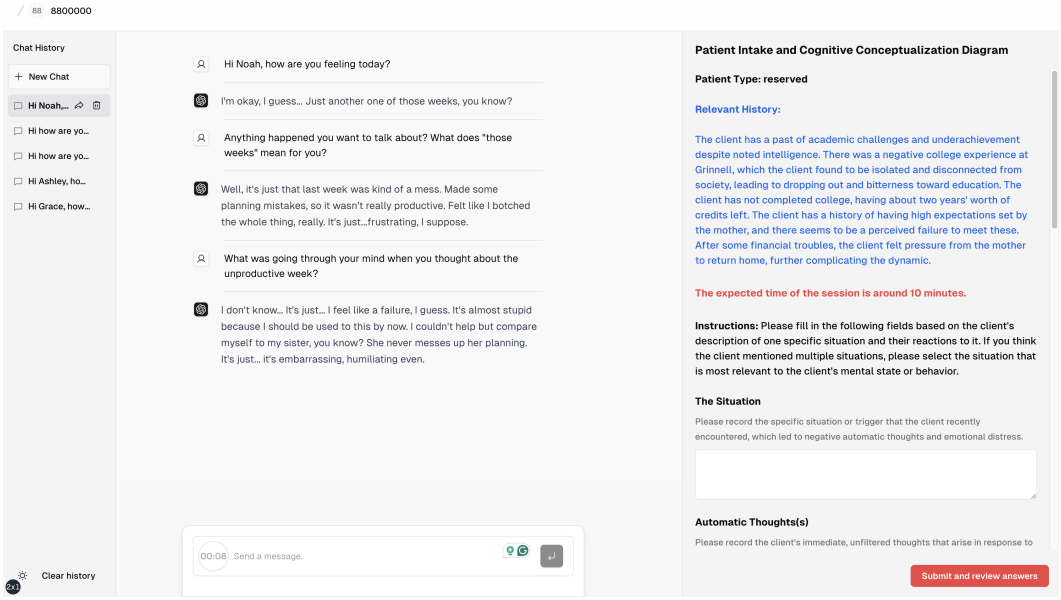


Figure 25: Our user interface of PATIENT-Ψ-TRAINER. Left: chatting window with PATIENT-Ψ; Right: forms to formulate the cognitive model (CCD). PATIENT-Ψ's relevant history and conversational style is shown to trainees at the onset of a session.

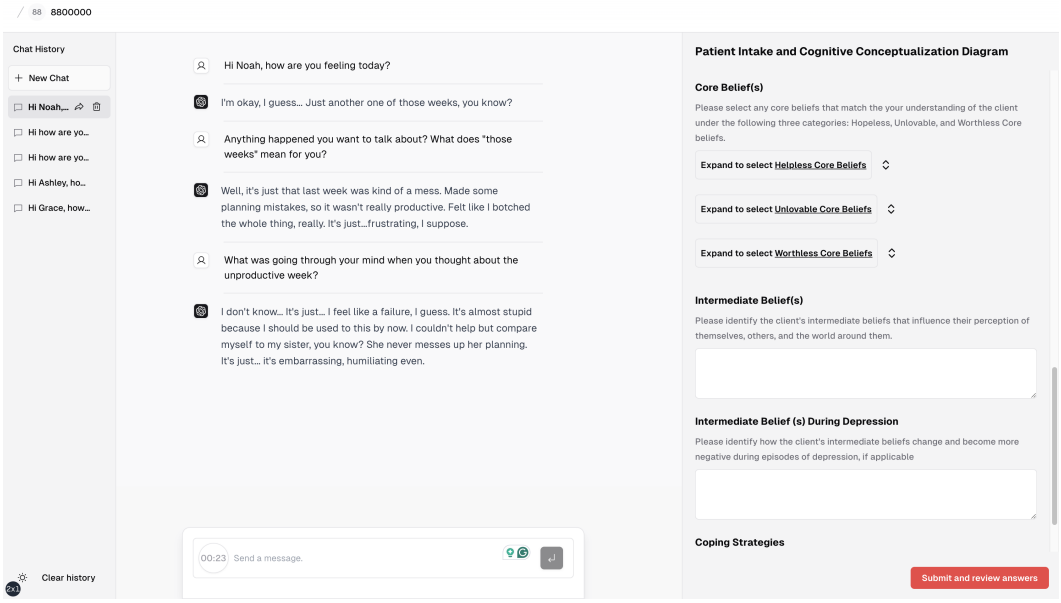


Figure 26: Our user interface of PATIENT-Ψ-TRAINER. Left: chatting window with PATIENT-Ψ; Right: forms to formulate the cognitive model (CCD). Trainees can scroll downwards to complete the components of the CCD in any order as they converse with PATIENT-Ψ.

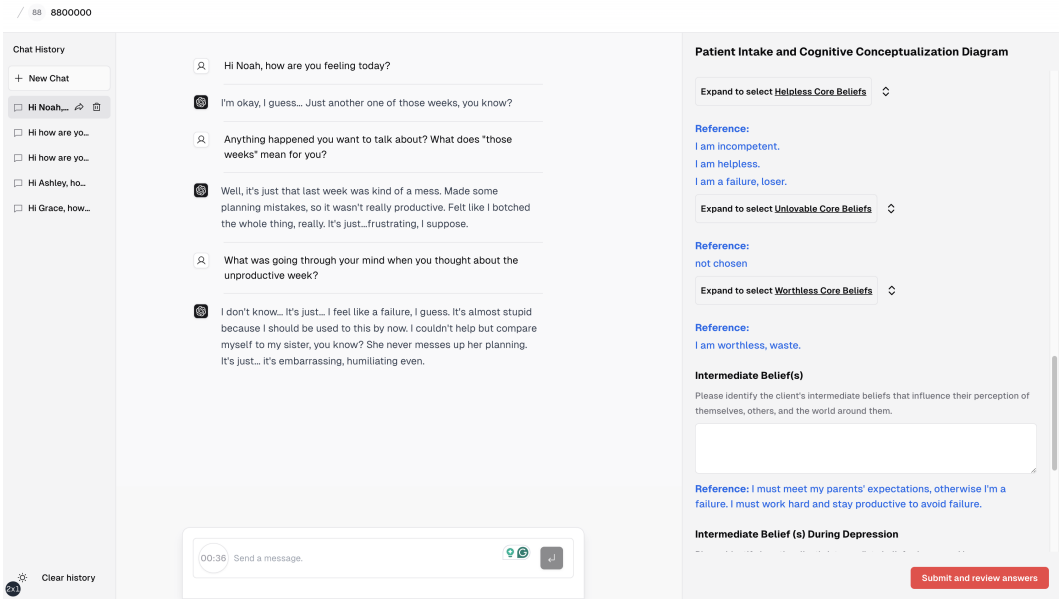


Figure 27: Our user interface of PATIENT-Ψ-TRAINER. Left: chatting window with PATIENT-Ψ; Right: forms to formulate the cognitive model (CCD). Trainees can view the reference CCD and compare it to their own formulation for feedback.