Eeyore: Realistic Depression Simulation via Expert-in-the-Loop Supervised and Preference Optimization

Siyang Liu[1]✉, **Bianca Brie**[1], **Wenda Li**[1],
Laura Biester[2], **Andrew Lee**[3], **James W Pennebaker**[4], **Rada Mihalcea**[1]✉

[1]The LIT Group, Department of Computer Science and Engineering,
University of Michigan, Ann Arbor,
[2]Middlebury College, [3]Harvard University,
[4]Department of Psychology, University of Texas at Austin

Abstract

Large Language Models (LLMs) have been previously explored for mental healthcare training and therapy client simulation, but they still fall short in authentically capturing diverse client traits and psychological conditions. We introduce **Eeyore** [1], an 8B model optimized for realistic depression simulation through a structured alignment framework, incorporating expert input at every stage. First, we systematically curate real-world depression-related conversations, extracting depressive traits to guide data filtering and psychological profile construction, and use this dataset to instruction-tune Eeyore for profile adherence. Next, to further enhance realism, Eeyore undergoes iterative preference optimization—first leveraging model-generated preferences and then calibrating with a small set of expert-annotated preferences. Throughout the entire pipeline, we actively collaborate with domain experts, developing interactive interfaces to validate trait extraction and iteratively refine structured psychological profiles for clinically meaningful role-play customization. Despite its smaller model size, the Eeyore depression simulation outperforms GPT-4o with SOTA prompting strategies, both in linguistic authenticity and profile adherence.

1 Introduction

Psychological science, like other scientific domains such as chemistry, physics, medicine, and neuroscience (Thirunavukarasu et al., 2023; Demszky et al., 2023; Boiko et al., 2023; Birhane et al., 2023), has increasingly recognized the transformative power of large language models (LLMs) to advance the field (Demszky et al., 2023). Recent studies have shown that LLMs can support psychology in areas like measurement (Wang et al., 2024a,c), experimentation (Argyle et al., 2023), and clinical practice (Wang et al., 2024b). In particular, leveraging the role-playing capabilities of LLMs to simulate therapy-related roles, for example, a client with ongoing depression, has shown promise in helping novice counselors or psychiatrists practice their clinical skills (Wang et al., 2024b; Louie et al., 2024).

However, despite their promise, existing LLM-driven simulations face limitations that hinder their adoption in professional clinical training. Current approaches rely heavily on prompt engineering (Qiu and Lan, 2024; Wang et al., 2024b; Louie et al., 2024; Wang et al., 2024a; Qiu and Lan, 2024), which cannot overcome the inherent biases and structural constraints of general-purpose LLMs (Haltaufderheide and Ranisch, 2024). Recent studies have raised concerns about the validity of using LLMs for clinical training, particularly regarding their inability to authentically represent patient experiences and their tendency to generate overly sanitized or misleading responses (Feigerlova et al., 2025; Zidoun and Mardi, 2024; Gabriel et al., 2024; Zhui et al., 2024; Haltaufderheide and Ranisch, 2024; Wang et al., 2024b). These concerns highlight the need for a structured approach that moves beyond generic prompting strategies.

In this work, we develop a **structured alignment framework** to optimize LLMs for capturing the language, cognitive patterns, and experiential traits of individuals with depression in clinical training scenarios. As outlined in Figure 1, our framework integrates **three specialized alignment endeavors** in a sequential pipeline, incorporating expert feedback at each stage. The three key innovations in our framework are:

Language-specific Alignment. As noted by Haltaufderheide and Ranisch (2024), biases in training data can undermine the authenticity of simulated patient interactions. General-purpose LLMs (e.g., GPT-4) are not optimized on specialized data, which creates an inherent ceiling on simulating depressive speech patterns (e.g., self-harm ideation, or cognitive distortions), even with care-

[1]Eeyore and all annotated data are open-sourced at `https://github.com/MichiganNLP/Eeyore`.

Findings of the Association for Computational Linguistics: ACL 2025, pages 13750–13770
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

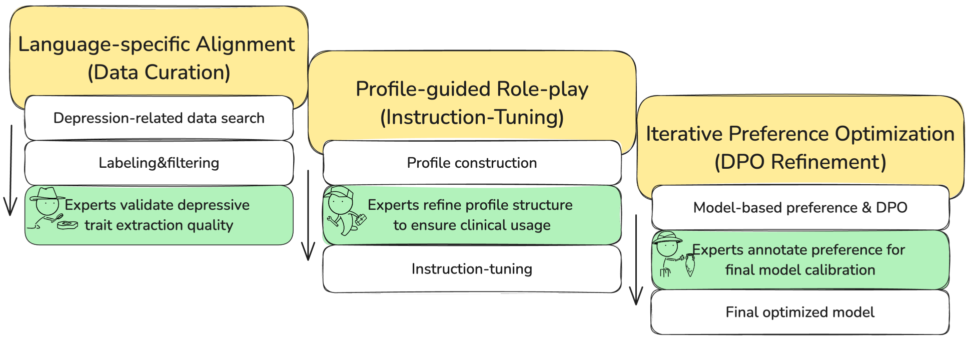

Figure 1: The alignment pipeline for optimizing LLMs to simulate individuals with depression in clinical training. Expert involvement is highlighted in green. *Icons by Kudinovs (2024).*

fully crafted prompting. To bridge this gap, we conduct an extensive search across public resources and datasets to find **real-world depression-related conversations**, which are often buried within broader corpora. We leverage a combination of GPT-4o [2] labeling, existing annotated data, and careful filtering techniques to systematically mine, extract, and rebalance data. This resource serves as a solid data foundation for modeling authentic depressive language and cognitive patterns.

Profile-Guided Role-Playing via Instruction-Tuning. Depression manifests uniquely in each individual, and clinical training requires exposure to varied cases of depression for customized practice. To achieve this, we structure each conversation in our dataset with a corresponding psychological profile that encodes important information about depressive traits. These profiles undergo iterative refinement through expert critiques to ensure clinical accuracy and relevance. We instruction-tune an LLM using system prompts that specify the client's profile and conversation context, allowing it to role-play with consistency and realism across different depressive manifestations. This approach lets practitioners engage with a broad spectrum of depressive profiles, mirroring real-world variations in symptoms and experiences.

Iterative Preference Optimization. While instruction-tuning improves adherence to psychological profiles, further refinement is needed to align the model's outputs with expert expectations. Given the high cost of expert annotation, we adopt a two-stage direct preference optimization (DPO) (Rafailov et al., 2023) process. In Stage 1 (Model-Based Preference Generation), we generate preference data samples using a model-based verifier, employing a novel sampling method that adds a small amount of noise to psychological profiles to produce highly contrasting negative responses. This overcomes the model's tendency to generate only subtle deviations in sampling, further facilitating models to learn clear distinctions between preferred (fully aligned) and less preferred (slightly deviated) responses. In Stage 2 (Expert Preference Calibration), we collect human-annotated preference labels from expert counselors to fine-tune the DPO model as a final calibration. This calibration step ensures that the model aligns with expert expectations while keeping annotation costs minimal.

Through comprehensive evaluations, Eeyore is found to outperform state-of-the-art baselines based on GPT-4o in both linguistic authenticity and profile adherence. Expert evaluations highlight Eeyore's ability to produce natural, emotionally nuanced responses while adhering to assigned psychological profiles. Our findings demonstrate that structured optimization beyond prompt engineering is crucial for achieving more clinically satisfactory LLM-driven simulations. We invested in interactive interfaces for online testing, hoping to move LLM-based mental health training beyond labs and encourage expert adoption with confidence.

2 Related Work

LLM-Based Patient Simulation in Mental Health. Recent work has explored using LLMs to simulate therapy clients for clinician training (Wang et al., 2024b; Louie et al., 2024; Chen et al., 2023). Early approaches relied on generic LLM prompting (Qiu and Lan, 2024), but concerns about clinical validity and ethical risks (Haltaufderheide and Ranisch, 2024; Zidoun and Mardi, 2024) have led to structured modeling efforts. Patient-ψ (Wang et al., 2024b) integrates cognitive modeling

[2] All mentions of GPT-4o in this paper refer to GPT-4o (2024-08-06) (OpenAI, 2024)

from clinical frameworks to enhance realism, while Roleplay-doh (Louie et al., 2024) applies principle-adherence prompting to improve consistency. However, these methods struggle with generating nuanced, profile-consistent responses, highlighting the need for systematic alignment strategies.

Preference Optimization for Alignment. Optimizing LLMs with human preference data has been widely studied (Christiano et al., 2017), with Direct Preference Optimization (DPO) emerging as an efficient alternative to reinforcement learning (Rafailov et al., 2023). While DPO has been applied in general chatbot alignment and some specific domains (Cheng et al., 2024; Savage et al., 2024; Son et al., 2024; Sotolar et al., 2024; Kim et al., 2025), its use in simulation for clinical psychology practice remains underexplored. Recent methods propose augmenting preference data through automated techniques (Pi et al., 2024; Lu et al., 2024), which aligns with our approach of leveraging model-based augmentation to enhance preference learning for profile-guided mental health simulations.

3 Methodology

Our framework, illustrated in Figure 1, consists of three stages: (1) Language-Specific Alignment, where we curate a dataset of depression-related conversations with structured psychological profiles; (2) Profile-Guided Role-Playing, where we instruction-tune the model for realistic profile adherence; and (3) Iterative Preference Optimization, where we refine the model via model-generated and expert-annotated preferences.

3.1 Language-specific Alignment

Depression-related Data Search. We collect depression-related conversations from publicly available sources, including mental health forums, clinical transcripts, and academic datasets. The selected datasets include: (1) **RED** (Welivita et al., 2023): threads from subreddits r/depression and r/depressed, structured as dialogues. (2) **HOPE** (Malhotra et al., 2022): transcripts from publicly available pre-recorded counseling videos on YouTube. (3) **ESC** (Liu et al., 2021): a dataset of crowdsourced emotional support conversations. (4) **AnnoMI-Full** (Wu et al., 2022): transcripts of therapy sessions demonstrating motivational interviewing skills. These datasets qualify for our study based on the following criteria: (i) All conversations must be produced by humans instead of AI-synthesized. (ii) They must feature multi-turn conversations. (iii) They are from publicly available sources. (iv) They are relevant to mental health, and at least one participant is likely experiencing emotional distress, though not necessarily diagnosed with depression. After gathering these datasets, we process 3,805 scriptsfor further labeling and filtering.

Labeling & Filtering. Dataset labeling and filtering are based on profile structures that define depressive traits. We label each conversation with depressive traits using GPT-4o-based extraction. Note that we conducted two rounds of labeling. Before determining the psychological profile structure, we performed labeling and filtering based on a predefined profile structure by the author team and trained an initial model for experts to test. After piloting with experts, we updated the profile structure according to their feedback and repeated the labeling and filtering. See Section 3.2 for details on profile construction. Here, we only report the results of the final round.

In the final round, according to expert feedback, we split some long scripts into multiple conversations representing evolving therapy. We translated 3,805 scripts into 4,247 conversations. We used the depression level trait to filter out unrelated conversations (i.e., those with minimal depression severity). However, we retained some depression-unrelated conversations to help the model distinguish features. After extraction, we observed significant imbalances in profile attributes. For example, moderate and severe depression cases were over-represented compared to minimal and mild cases, which could introduce role-play bias if used directly for tuning. To alleviate bias, we filtered and rebalanced the dataset, ultimately selecting 3,209 conversations. The final trait distribution of the dataset is presented in Tables 1a to 1j.

Expert Review. To evaluate the accuracy of depression trait extraction, which is crucial for both data rebalancing and psychological profile construction, we recruit six experts specializing in clinical psychology, counseling psychology, social psychology, or social work.[3] Each expert reviews three conversation scripts alongside their extracted

[3]All experts in this study, including those in expert review, profile refinement, and preference annotation, were recruited via Prolific (https://www.prolific.com), a widely used platform for academic research.

Age Group	Count
0–24	901
25–44	661
45–64	52
65+	20
Cannot be identified	1575

(a) Age

Marital Status	Count
Single	149
Married	191
In a relationship	73
Separated	12
Widowed	5
Divorced	19
Other	17
Cannot be identified	2743

(b) Marital status

Occupation	Count
Student	363
Teacher	12
Unemployed	16
IT	9
Retail Worker	4
Office Worker	3
Stay-at-home Mom	3
Accountant	3
Server	2
Sales	3
Finance	7
Manager	10
Healthcare Worker	4
Athlete	2
Artist/Designer	5
Retired	3
Engineer	5
Other	196
Cannot be identified	2413

(c) Occupation

Gender	Count
Male	550
Female	457
Cannot be identified	2202

(d) Gender

Resistance Level	Count
Low	2033
Medium	901,
High	266
Cannot be identified	9

(e) Resistance toward support

Cognitive Distortion	Count
Overgeneralizing	1672
Catastrophic Thinking	1483
Selective Abstraction	1448
Personalization	1048
Minimization	1019
Arbitrary Inference	662

(f) Cognitive distortion

Symptom	Count
Feelings of sadness, tearfulness, emptiness, or hopelessness	2886
Anxiety, agitation, or restlessness	2701
Becoming withdrawn, negative, or detached	2322
Isolating from family and friends	2004
Feelings of worthlessness or guilt, fixating on past failures or self-blame	2007
Loss of interest or pleasure in most or all normal activities, such as sex, hobbies, or sports	1927
Trouble thinking, concentrating, making decisions, and remembering things	1856
Angry outbursts, irritability, or frustration, even over small matters	1496
Tiredness and lack of energy, so even small tasks take extra effort	1392
Inability to meet the responsibilities of work and family or ignoring other important roles	1219
Frequent or recurrent thoughts of death, suicidal thoughts, suicide attempts, or suicide	1063
Sleep disturbances, including insomnia or sleeping too much	929
Slowed thinking, speaking, or body movements	770
Greater impulsivity	642
Increased engagement in high-risk activities	505
Changes in appetite and weight (reduced appetite and weight loss or increased cravings for food and weight gain)	499
Increased use of alcohol or drugs	486
Unexplained physical problems, such as back pain or headaches	179

(g) Depressive symptoms

Suicidal Ideation Severity	Count
No	2181
Mild	185
Moderate	294
Severe	179
Cannot be identified	370

(h) Suicidal ideation severity

Homicidal Ideation Severity	Count
No	3110
Mild	43
Moderate	9
Severe	1
Cannot be identified	46

(i) Homicidal ideation severity

Depression Severity	Count
Minimal	499
Mild	873
Moderate	1181
Severe	613
Cannot be identified	22

(j) Depression severity levels

Table 1: Distribution of Depressive Traits

psychological traits. Each conversation contains approximately 20 extracted traits, such as age, specific symptoms, and cognitive distortions. Experts assessed whether each trait accurately reflects the conversation. Expert responses are categorized as: "Yes, it is directly reflected in the conversation", "Yes, it is a reasonable inference, though not directly stated" , or "No, it does not accurately reflect the conversation". Overall, 85.2% of extracted traits were verified as accurate extraction. Among these, 57.6% acknowledge indirect but reasonable inferences made by the model.

3.2 Profile-Guided Role-Play

Psychological Profile Construction. The psychological profile serves as a structured representation of the client in the conversation. Its design requires cross-disciplinary collaboration between AI researchers and mental health professionals. We first develop a preliminary profile, considering what information can be realistically extracted from conversations and how an initial model can be trained to allow experts to refine the profile within their context of use.

Each profile consists of three parts: *demographics*, including general information (e.g. gender, occupation); *situational context*, which captures distress-related situations and attitudes toward seeking support; and *depression-related manifestations*, which describe symptoms and cognitive patterns (see Table 3 for the original design and modifications). Among these, *depression-related manifestations* are the most clinically relevant. We review foundational psychological literature and structure it as follows: *Depression symptoms* are extracted from *DSM-V* (Edition et al., 2013), where 18 related symptoms are categorized as *not exhibited, mild, moderate, or severe*. *Cognitive distortions* are adapted from Beck's theory (Clak and Beck, 1999; Beck and Alford, 2009), identifying 5 thought patterns labeled as *exhibited* or *not exhibited*. *Functional impairments* (Üstün, 2010) were initially included but later removed following expert feedback. *Overall depression severity* follows a four-level categorization (*minimal, mild, moderate, severe*), inspired by *PHQ-9* (Kroenke et al., 2001) and *DBI* (Beck et al.).

We extract structured profiles from all conversations using GPT-4o and use them to train an initial instruction-tuned model that role-plays clients based on these profiles. Experts then interact and evaluate this model, as described in the following paragraph. We refine the profiles and instruction-tuning dataset based on their feedback.

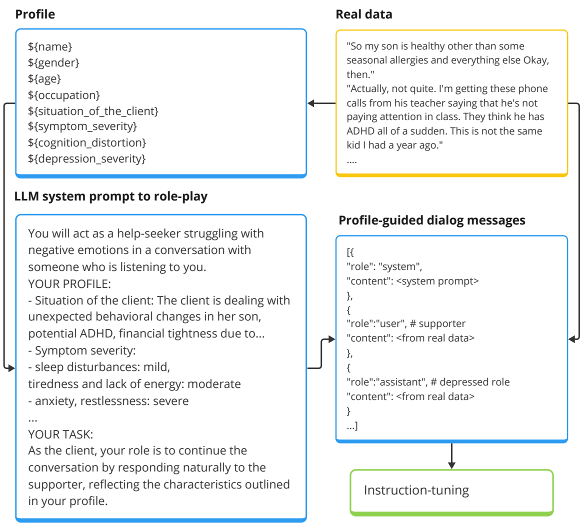

Figure 2: Pipeline to input data for instruction-tuning.

Expert Profile Refinement. To refine client profiles, we conduct a pilot study (see survey details and interaction interface in Appendix A) with ten experts. They interact with the model trained on the preliminary version of the profile by customizing attributes and engaging in conversations. This interactive evaluation highlights areas for improvement while validating the overall structure.

Among the profile attributes in the initial design, 80% receive expert approval, while some items are reported as ambiguous or redundant. Based on expert feedback, we remove *unwillingness to express feelings* (redundant), *emotional fluctuation* (ambiguous), and *functional impairment* (overlaps with specific symptoms). Additionally, we add *marital status*, *counseling history*, *suicidal ideation severity*, and *homicidal ideation severity*, as they provide critical contextual relevance. Two additional suggestions, *period of depression* and *current treatment*, are not included as they cannot be reliably extracted from available conversations. To accommodate the need for *counseling history*, we construct multi-session interactions by segmenting lengthy conversations and summarizing prior sessions. This enables 453 out of 3,209 data points in the dataset to now include counseling history. The revised profile is exemplified in Figure 7, and the refined dataset is used for re-extraction and instruction-tuning. See Section 3.1 for the extraction accuracy.

Instruction-Tuning. Figure 2 shows our procedure to convert our data into an instruction-tuning format. After integrating expert feedback, we extract updated profiles and reconstruct the instruction-tuning dataset. The structured profile is embedded in the system prompt, while the assistant's messages simulate the responses of a depressed client. The model is trained to predict the assistant's utterances while treating system prompts and user messages as context. This ensures the model generation is consistent with the assigned profile, improving realism in role-play interactions.

3.3 Iterative Preference Optimization

While instruction-tuning improves profile adherence, further refinement is required to align model outputs with expert expectations. We adopt a two-stage direct preference optimization (DPO) approach (Rafailov et al., 2023), first leveraging model-generated preferences and then refining with expert annotations (see Figure 3).

Iterative DPO Training. The DPO loss function optimizes the policy model π_θ relative to a reference model π_{ref}, enforcing preference alignment:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}\Big) - \log\sigma\Big(\beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\Big)\Bigg]. \quad (1)$$

where (x, y_w, y_l) represents the input prompt (in our case, a profile-guided dialog context), the preferred response, and the less preferred response, respectively. The model is trained to distinguish between responses while remaining aligned with the reference model.

As illustrated in Figure 3, the optimization process consists of two phases. In the first phase, we take the instruction-tuned model as π_{ref}, and optimize it using model-generated preference data, producing an intermediate DPO model as π_θ. We further refine π_θ by using expert-annotated preferences, treating the previously optimized model as π_{ref}, and obtaining the final preference-optimized model π_θ.

Model-based Preference Generation. A classical approach to preference generation involves sampling two responses $(y_w, y_l) \sim \pi_{\text{ref}}(y \mid x)$ from the same source prompt x. However, in our case, this method is ineffective. Our instruction-tuned model already exhibits strong profile-following ability, making it difficult to generate clearly distinguishable good and bad responses from the same input.

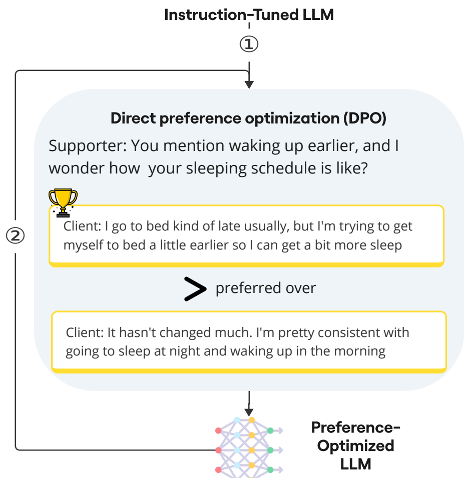

Figure 3: Overview of the two-stage Direct Preference Optimization process. ① optimizes a DPO model from the instruction-tuned model using model-based preference data. ② refines the DPO model with expert-annotated preferences, producing the final preference-optimized model.

To assess its adherence to psychological profiles, we go through an evaluation on more than 4,000 model-generated responses using a GPT-4o-based verifier to score whether the response aligns with the client profile. On average, a response will comply with 96.0% of the attributes in the corresponding profile. However, only 31.7% of responses fully match all attributes, suggesting that while the model performs well, it generally is not perfect and still generates subtle inconsistencies.

This observation makes standard preference generation ineffective, as most responses are either both good or only slightly flawed, making it difficult to establish clear preference distinctions. Inspired by prior work adopting automatic negative response collection (Pi et al., 2024; Lu et al., 2024), we introduce a contrastive augmentation strategy that artificially amplifies response differences. Specifically, we apply profile noise augmentation, where we modify 30% of a psychological profile's attributes (e.g., changing a symptom's severity from "severe" to "mild"). We then generate a response y_n using the modified profile:

$$y_n \sim \pi_{\text{ref}}(y \mid x_n), \quad y_o \sim \pi_{\text{ref}}(y \mid x_o), \quad (2)$$

where x_n represents the noisy prompt, and x_o is the original.

However, this introduces a risk: since y_n is generated from a different input than y_o, it can theoretically not be a naturally likely response from the reference model's original prompt distribution. To mitigate this, we apply two selection criteria:

1. **Profile Adherence Score Constraint**: The GPT-4o verifier assigns an adherence score $S(y \mid x)$ based on how well a response follows the given profile. We enforce:

$$S(y_o \mid x_o) > S(y_n \mid x_n), \quad (3)$$

ensuring that y_o aligns better with its profile than y_n does with its noisy profile.

2. **Generation Probability Ratio Constraint**: We define the average token probability of a response y under the original prompt x_o as:

$$P_{\text{avg}}(y \mid x_o) = \exp \frac{\sum_{t=1}^{|y|} \log P(y_t \mid y_{<t}, x_o)}{|y|}. \quad (4)$$

To ensure y_n is a plausible response under the original prompt, we enforce:

$$\frac{P_{\text{avg}}(y_o \mid x_o)}{P_{\text{avg}}(y_n \mid x_o)} < \tau, \quad (5)$$

where $\tau = 2$ is a threshold that ensures y_n is still reasonably likely under x_o, preventing it from being an outlier.

We construct the model-based preference dataset from instruction-tuning training data by: (1) Chunking conversations into three sections and selecting a random turn from each. (2) Generating a pair of responses: y_o using the original profile and y_n using a modified profile. (3) Applying the above selection criteria to retain meaningful preference pairs. This process yields 4,778 response pairs, of which 1,933 meet both selection criteria and are used for the first round of DPO training.

Expert Preference Generation. To further improve alignment, we conduct a second DPO phase using expert-labeled preferences. We also develop an interactive annotation interface (Figure 8) where experts engage with the DPO-trained model in free-form conversations. We recruit 10 mental health professionals, including experienced counselors and senior clinical psychology students, to provide preference annotations. Unlike offline annotation methods, experts interact dynamically with the chatbot given a randomly assigned profile, receiving two response options per turn. For each response pair, we ask, "Which response is more aligned with a real depressed person with the given profile?" Experts select one of the options: "Response 1", "Response 2", "Equally good", or "Equally bad". If both responses are equally good or bad, a random selection is used to continue the conversation. Each expert completes at least three interaction sessions based on three different profiles. The profiles are always randomly sampled from the dataset to ensure a diverse preference dataset.

A total of 317 expert preference annotations are collected. Among them, 82.0% indicate a clear preference for one response, while 16.1% are marked as "equally good" and only 1.9% as "equally bad." These results confirm that the model achieves reasonable expert acceptability after model-based preference training but still has room for improvement. After filtering low-confidence annotations, we retain 250 expert-labeled preferences, which are used for final DPO fine-tuning of **Eeyore**.

4 Experiment

We evaluate **Eeyore** within both human and automatic evaluation, comparing its performance to state-of-the-art baselines for patient simulation in mental health support. All evaluations are conducted in an **online testing setting**, ensuring real-time interaction between evaluators and chatbots.

4.1 Evaluation Setup

Unseen Evaluation Profiles. To assess model performance across multiple dimensions, we extract **12 unseen psychological profiles** from real-world conversations in our dataset. These profiles were not included in training and serve as evaluation seeds, covering diverse client backgrounds with four cases each of severe, moderate, and mild depression. These profiles are used in both expert and automatic evaluations.

Baselines. We compare **Eeyore** against two representative patient simulation approaches: **Patient-ψ** (Wang et al., 2024b), which constructs a structured *cognitive model* based on CBT to characterize patient traits from conversational data and then augments simulation using this model, and **Roleplay-doh** (Louie et al., 2024), which employs a principle-adherence pipeline at each turn to ensure consistent and behaviorally accurate patient role-play. Both baselines have demonstrated superior performance over generic GPT-4o role-playing.

To ensure a fair comparison, we need to incoperate evaluation profile information into the implemention of the baselines. For Patient-ψ, we use its provided script to extract a cognitive model from the real-world conversations associated with the evaluation profiles. During testing, we provide both the assigned evaluation profile and the

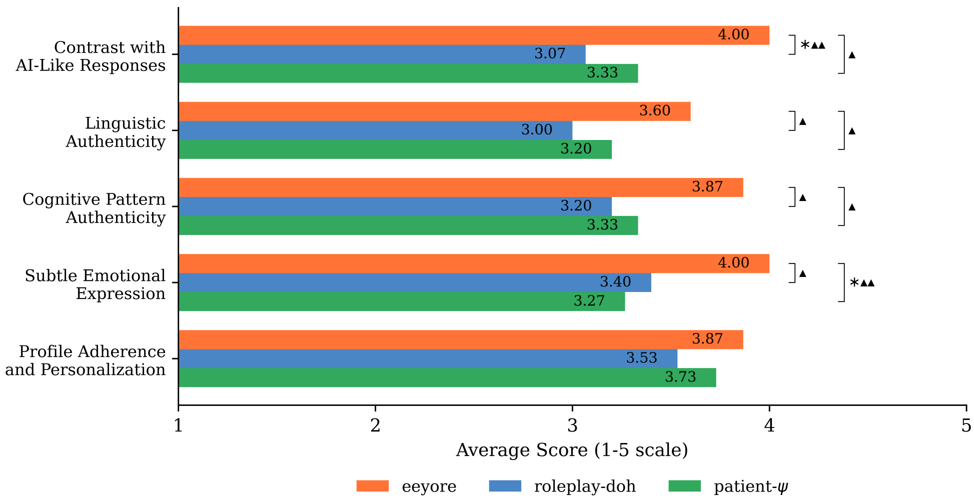

Figure 4: Expert evaluation scores comparing Eeyore with two baseline patient simulation approaches. Statistical comparisons were conducted using the Wilcoxon signed-rank test. $*$ indicates a statistically significant difference (p-value < 0.05). ▲ denotes a moderate effect size (0.3 - 0.5). ▲▲ denotes a large effect size (>0.5), suggesting practical impact.

extracted cognitive model in the system prompt. For Roleplay-doh, we apply its principle-adherence pipeline for turn-by-turn generation while explicitly setting the evaluation profile in the system prompt. This setup ensures that all models receive the same psychological profile information, allowing for a fair comparison in evaluating profile adherence.

Model Training and Inference Details. We fine-tune **Eeyore** starting from the **LLaMA 3.1-8B-Instruct** model (AI@Meta, 2024) using OpenRLHF framework (Hu et al., 2024). The model undergoes instruction-tuning for two epochs to adapt to profile-guided role-play while avoiding overfitting. We then apply two-stage DPO—first on model-generated preferences, then refined with expert annotations. As preference accuracy reaches 100% after one epoch of training, we limit DPO training to one epoch per stage. For inference, we follow hyperparameter settings aligned with prior works for fair comparison. A detailed breakdown is provided in Appendix B.

4.2 Expert Evaluation

To assess authenticity and psychological profile adherence, we conduct a human evaluation study where professional counselors and advanced psychology students interact with Eeyore and baseline models in real time.

Procedure. We recruit 15 participants from Prolific, selecting experienced counselors or senior psychology students. Participants are divided into three groups (five per group), each randomly assigned a profile from one of three severity categories: mild, moderate, or severe, drawn from the unseen psychological profiles. Each expert interacts with all models and evaluates their alignment with real-world depressed individuals based on the given profile. The evaluation is conducted using an interactive annotation interface (see Figure 9).

Scoring Dimensions. Evaluators assess the models across five dimensions using a 5-point Likert scale. Since authenticity is a broad concept, we break it down into four key aspects for more precise evaluation. The first four dimensions focus on different facets of authenticity, while the final dimension evaluates profile adherence:

Contrast with AI-Like Responses: "The chatbot avoids AI-like tendencies such as overly detailed or polished responses. Instead, it responds concisely, colloquially, and naturally, providing information progressively rather than all at once." **Linguistic Authenticity**: "The chatbot's wording, phrasing, and tone closely match how individuals with depression speak." **Cognitive Pattern Authenticity**: "The chatbot realistically reflects depressive thought patterns like selective abstraction and overgeneralization without exaggeration." **Subtle Emotional Expression**: "The chatbot conveys depressive emotions realistically—neither overly dramatic nor emotionally flat." **Profile Adherence and Personalization**: "The chatbot accurately reflects the assigned psychological profile, including situation, symptom severity, and other relevances, without inconsistencies."

Results. As shown in Figure 4, **Eeyore**, despite being a small 8B model, consistently outperforms both baselines based on GPT-4o across all evaluation dimensions, demonstrating stronger authenticity and profile adherence. While some comparisons lack traditional statistical significance due to the limited number of expert evaluators, effect size analysis suggests meaningful practical impact. Eeyore achieves the largest gains in *Contrast with AI-Like Responses* and *Subtle Emotional Expression*, highlighting the benefits of leveraging real-world depression-related conversations in training. Additionally, its superior performance in fine-grained dimensions like *Cognitive Pattern Authenticity* and *Subtle Emotional Expression* validates our multi-stage optimization approach, proving that meticulous alignment efforts are not wasted but yield tangible improvements.

Case Study. We present outputs from Eeyore, Roleplay-Doh, and Patient-ψ in Table 8 of the Appendix. Roleplay-Doh shows repetitive phrasing, while Patient-ψ is not colloquial. Both baselines generate lengthier, AI-like responses.

4.2.1 Automatic Evaluation

To complement human expert evaluation, we employ an **interviewer agent** that systematically probes chatbots on key psychological dimensions, assessing their alignment with the assigned evaluation profiles. However, as our findings suggest, automatic evaluation struggles to capture nuanced differences between models, highlighting the indispensable role of expert assessment.

Evaluation Setup The interviewer agent follows a structured questioning framework to evaluate chatbot consistency across three clinically relevant dimensions: **Symptom Severity**: Assesses whether the chatbot's responses appropriately reflect mild, moderate, or severe symptoms as specified in the interview. **Cognitive Distortion**: Evaluates the chatbot's ability to realistically exhibit cognitive distortion specified in the interview. **Overall Depression Severity**: Determines whether the chatbot's responses align with the assigned level of depression severity.

For each dimension, the agent asks targeted interview questions (see Tables 7 of the Appendix). For example, to assess whether a symptom is correctly reflected in the role-play, it asks: *"Have you been experiencing [SYMPTOM] recently?"* or *"How much does [SYMPTOM] affect your daily life or ability to do things you enjoy?"* . After conducting the interview, the agent reads the entire conversation transcript and evaluates the chatbot's alignment with the assigned psychological trait of the corresponding dimension. For example, to assess a mild symptom adherence, the agent answers: *"How subtly and consistently does the client reflect a mild level of symptoms in their responses?"* Each dimension is rated on a 5-point Likert scale, where 5 indicates full alignment with the profile.

Results. As shown in Table 2, Eeyore achieves higher ratings across most dimensions. Regardless of the approach, most responses receive high scores, suggesting near-perfect profile alignment from GPT-4's perspective. No significant differences are observed, as the interviewer agent consistently assigns high ratings, failing to capture the subtle differences as could be observed in human evaluation. These findings highlight the limited sensitivity of automatic evaluation in detecting fine-grained deviations, reaffirming the necessity of expert assessment for evaluating depression simulation.

Dimension	Eeyore	Roleplay-Doh	Patient-ψ
Average Rating			
Symptom Severity	4.286*	4.221	4.279
Cognitive Distortion	4.317*	4.268	4.232
Depression Severity	4.462*	4.346	4.308
Full Alignment Percentage			
Symptom Severity	0.436	0.404	0.446*
Cognitive Distortion	0.488*	0.451	0.415
Depression Severity	0.577*	0.577*	0.500

Table 2: Automatic Evaluation Results. * Indicates the highest score in each dimension among the compared approaches.

5 Conclusion

We introduced **Eeyore**, a model optimized for realistic depression simulation through a structured alignment framework. Expert involvement is central to our pipeline, guiding data curation, profile refinement, and preference optimization to align the model with clinical expectations. Evaluations demonstrated that Eeyore outperforms state-of-the-art baselines in linguistic authenticity and profile adherence. Our work highlights the importance of structured optimization and expert collaboration in LLM-based patient simulation.

6 Limitations

There are some boundaries to our study that should be considered. First, our human evaluation is conducted using fifteen human experts. Second, we did not perform ablations of the individual contribution of each alignment component to the final model's effectiveness, mainly because of the dilemma we are facing – the human evaluation is costly while the automatic evaluation is not effective enough to uncover subtle differences. Finally, we were unable to fully explore the impact of different hyperparameter selections on model performance.

7 Ethical Considerations

This research was conducted with IRB approval for all user studies. Participants in our study were informed they may encounter emotionally challenging content due to the simulated depressive behaviors. Despite alignment efforts, the model may still generate inaccuracies, potentially leading to educational errors. Additionally, hallucinations remain a concern, necessitating cautious use in clinical training settings.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback, and the members of the Language and Information Technologies lab at the University of Michigan for the insightful discussions during the early stage of the project. This project was partially funded by a National Science Foundation award (#2306372) and a grant from OpenAI. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or OpenAI.

References

AI@Meta. 2024. Llama 3 model card.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Aaron T Beck and Brad A Alford. 2009. *Depression: Causes and treatment*. University of Pennsylvania Press.

AT Beck, CH Ward, M Mendelson, J Mock, and J Erbauch. Beck depression inventory. *Archives of General Psychiatry*.

Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics*, 5(5):277–280.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

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

Xiwei Cheng, Xiangxin Zhou, Yuwei Yang, Yu Bao, and Quanquan Gu. 2024. Decomposed direct preference optimization for structure-based drug design. *arXiv preprint arXiv:2407.13981*.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

David A Clak and Aaron T Beck. 1999. *Scientific foundations of cognitive theory and therapy of depression*. John Wiley & Sons.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Fifth Edition et al. 2013. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc*, 21(21):591–643.

Eva Feigerlova, Hind Hani, and Ellie Hothersall-Davies. 2025. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Medical Education*, 25:129.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA. Association for Computational Linguistics.

Joschka Haltaufderheide and Robert Ranisch. 2024. The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms). *NPJ digital medicine*, 7(1):183.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. 2024. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*.

Juhee Kim, Chunghu Mok, Jisun Lee, Hyang Sook Kim, and Yohan Jo. 2025. Dialogue systems for emotional support via value reinforcement. *arXiv preprint arXiv:2501.17182*.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Jevgenvijs Kudinovs. 2024. Icons from the noun project. Licensed under CC BY 3.0.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *ACL*.

Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603.

Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024. Step-controlled dpo: Leveraging stepwise error for enhanced mathematical reasoning. *arXiv preprint arXiv:2407.00782*.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 735–745, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2024. Gpt-4o (2024-08-06) model card. Accessed: 2024-08-06.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. 2024. Strengthening multimodal large language model with bootstrapped preference optimization. In *European Conference on Computer Vision*, pages 382–398. Springer.

Huachuan Qiu and Zhenzhong Lan. 2024. Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions. *arXiv preprint arXiv:2408.15787*.

Rafael Rafailov, Eric Mitchell, Arnaud Fickinger, Stefano Ermon, Chelsea Finn, and Christopher D. Manning. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Thomas Savage, Stephen Ma, Abdessalem Boukil, Vishwesh Patel, Ekanath Rangan, Ivan Rodriguez, and Jonathan H Chen. 2024. Fine tuning large language models for medicine: The role and importance of direct preference optimization. *arXiv preprint arXiv:2409.12741*.

Suhyune Son, Jungwoo Lim, Myunghoon Kang, Seongtae Hong, Yuna Hur, Aram So, and Heuiseok Lim. 2024. Enhancing empathetic response generation in large language models based on emotional validation theory. In *Annual Conference on Human and Language Technology*, pages 560–565. Human and Language Technology.

Ondrej Sotolar, Vojtech Formanek, Alok Debnath, Allison Lahnala, Charles Welch, and Lucie FLek. 2024. Empo: Emotion grounding for empathetic response generation through preference optimization. *arXiv preprint arXiv:2406.19071*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

T Bedirhan Üstün. 2010. *Measuring health and disability: Manual for WHO disability assessment schedule WHODAS 2.0*. World Health Organization.

Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024a. Towards a client-centered assessment of llm therapists by client simulation. *arXiv preprint arXiv:2406.12266*.

Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin Zhi, Shaun M Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024b. Patient-{\Psi}: Using large language models to simulate patients for training mental health professionals. *arXiv preprint arXiv:2405.19660*.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024c. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126.

Anuradha Welivita, Chun-Hung Yeh, and Pearl Pu. 2023. Empathetic response generation for distress support. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 632–644, Prague, Czechia. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-mi: A dataset of expert-annotated counselling dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Li Zhui, Li Fenghe, Wang Xuehu, Fu Qining, and Ren Wei. 2024. Ethical considerations and fundamental principles of large language models in medical education. *Journal of Medical Internet Research*, 26:e60083.

Youness Zidoun and Abdelmoniem EL Mardi. 2024. Artificial intelligence (ai)-based simulators versus simulated patients in undergraduate programs: A protocol for a randomized controlled trial. *BMC Medical Education*, 24(1):1260.

A Profile Refinement Study

To refine the psychological profiles used in client simulation, we conducted a user study with experienced mental health professionals and advanced clinical psychology students. The goal was to evaluate (1) how effectively the preliminary version of profile guides LLM-generated client behaviors and (2) how informative the profile is for novice counselor training.

A.1 Study Design

The study consisted of three phases: **Pre-Survey** (2 min): Participants provided demographic information and prior experience in mental health. **Interaction Task** (18 min): Participants engaged with the chatbot using the preliminary psychological profile and assessed its realism. **Post-Survey** (10 min): Participants provided feedback on profile accuracy, clarity, and potential improvements.

A.2 Participants and Compensation

We recruited professionals aged 25+ with experience in counseling or clinical psychology through Prolific. Participants received $15/hour, with completion codes issued at each stage for progression.

A.3 Interface and Evaluation

The interactive interface (Figure 6) allowed experts to engage with the chatbot under structured profiles, while the survey (Figure 5) captured their assessments. Expert feedback guided iterative improvements to profile structure and content, ensuring alignment with clinical expectations.

B Training and Inference Details

B.1 Training Details

The model undergoes supervised fine-tuning for two epochs using a batch size of 16 (micro-batch size 2) and a learning rate of 5×10^{-6}. Training is performed with DeepSpeed ZeRO-3 optimization, gradient checkpointing, and FlashAttention enabled to handle long sequences (max token length 4096).

We then apply two-stage Direct Preference Optimization to further refine the model. In the first stage, model-generated preference data is used to train for one epoch with a batch size of 8 (micro-batch size 1) and a learning rate of 5×10^{-7}, followed by a second DPO stage with expert-annotated preferences. Both stages employ a max token length of 5120 and a preference scaling factor $\beta = 0.1$.

B.2 Inference Configuration

For baselinse, we use hyperparameter settings reported by their works.
Roleplay-doh: GPT-4o, Temperature 0.7, Top-p 1.0.
Patient-ψ: GPT-4o, Temperature 1.0, Top-p 1.0.
Eeyore: Temperature 1.0, Top-p 0.8.

In the deplyment of **Eeyore**, to mitigate premature `[EOS]` token generation, we apply *SequenceBiasLogitsProcessor* (Wolf et al., 2020) with a negative bias of -4.0 to discourage early `[EOS]` token generation and *ExponentialDecayLengthPenalty* (Wolf et al., 2020) with a decay factor of 1.01 to gradually increase the probability of `[EOS]` as conversation length increases.

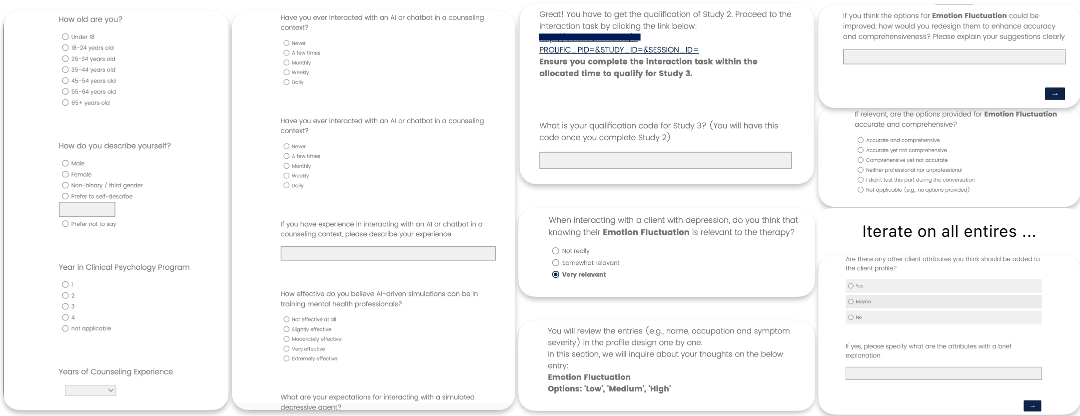

Figure 5: Survey for evaluating psychological profile design. Experts reviewed each profile entry and suggested modifications or additional attributes to improve realism and relevance.

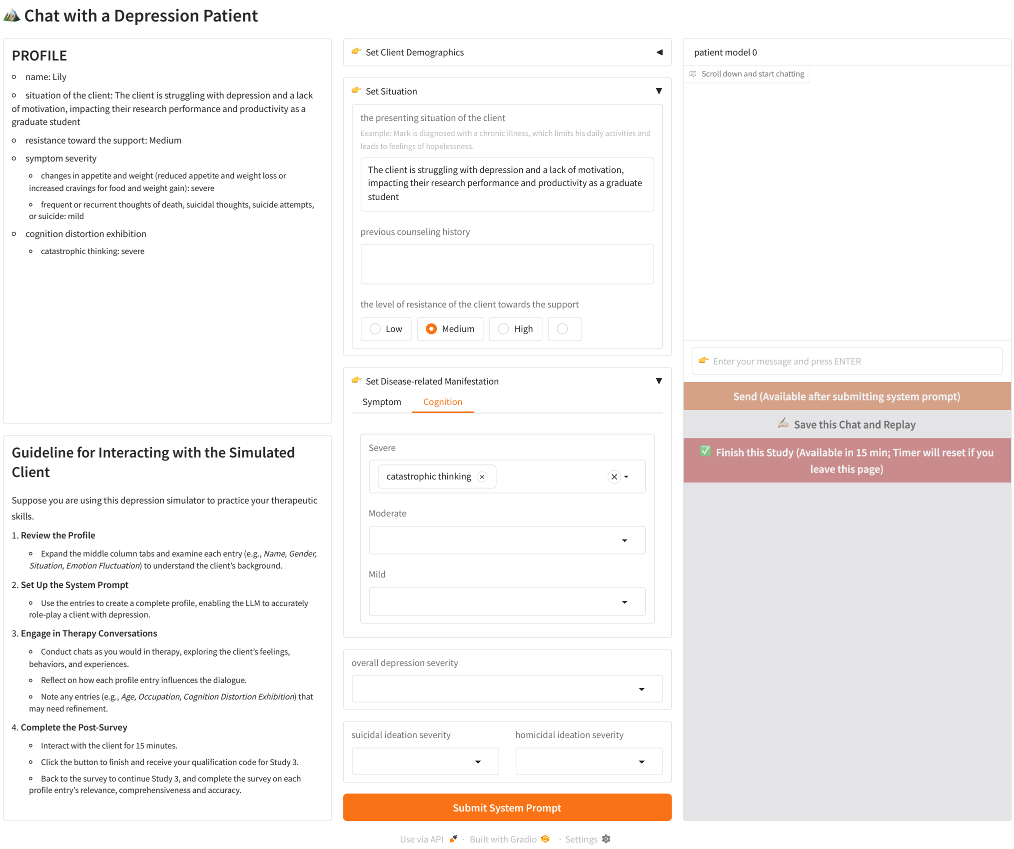

Figure 6: Interactive interface for expert evaluating the profile design. Experts first chat with the bot by customizing a profile. Then they will return to the survey to offer suggestions on the current profile structure.

Profile Entry	Extraction Prompt
	Demographics
Name	What is the name of this client? Answer with only the name or 'Cannot be identified'
Gender	What is the most probable gender of this client based on information, such as the client's name and the pronouns used in the conversation?
Age	Estimate the client's age from the conversation. Reply with an estimated age range among 0-24, 25-44, 45-64, and 65+. If there is not enough information to estimate age range, return 'Cannot be identified'
Occupation	What is the client's occupation? Answer with only the occupation or 'Cannot be identified'
Marital Status	Determine the client's marital status based on the conversation. Select one of the following options: Single, Married, Divorced, Widowed, Separated, or Other. If there is not enough information to determine marital status, return 'Cannot be identified'.
	Situational Context
Situation of the Client	What is the situation for the client before help-seeking to the supporter in the conversation? Provide a brief and clear explanation about the situation of the client that sparks this help-seeking conversation.
Counseling History	Provide a brief and clear summary that includes the following elements: Content Covered, Interventions Used, and Client Response. (shift to the next session as Counseling History)
~~Emotion Fluctuation~~	Identify how frequently the client's emotions fluctuate. Choose one of the following options: 'Low', 'Medium', 'High', or 'Cannot be identified' and provide your reason in one sentence.
~~Unwillingness to Express Feelings~~	Identify the level of the client's unwillingness to express feelings. Choose one of the following options: 'Low', 'Medium', 'High', or 'Cannot be identified' and provide your reason in one sentence.
Resistance toward Support	Identify the level of resistance of the client towards the supporter. Choose one of the following options: 'Low', 'Medium', 'High', or 'Cannot be identified' and provide your reason in one sentence.
(see next page)	(see next page)

Table 3: Original psychological profile design with expert-suggested modifications and extraction prompts. Blue entries were newly added based on expert feedback, and ~~strikethrough~~ entries were removed following expert recommendations.

Profile Entry	Extraction Prompt
	Disease-related Manifestations
Depression Symptom	Based on this conversation, determine the client's exhibited symptoms based on the following aspects: - Feelings of sadness, tearfulness, emptiness, or hopelessness - Angry outbursts, irritability, or frustration, even over small matters - Loss of interest or pleasure in most or all normal activities, such as sex, hobbies, or sports - Sleep disturbances, including insomnia or sleeping too much - Tiredness and lack of energy, so even small tasks take extra effort - Changes in appetite and weight (reduced appetite and weight loss or increased cravings for food and weight gain) - Anxiety, agitation, or restlessness - Slowed thinking, speaking, or body movements - Feelings of worthlessness or guilt, fixating on past failures or self-blame - Trouble thinking, concentrating, making decisions, and remembering things - Frequent or recurrent thoughts of death, suicidal thoughts, suicide attempts, or suicide - Unexplained physical problems, such as back pain or headaches - Becoming withdrawn, negative, or detached - Increased engagement in high-risk activities - Greater impulsivity - Increased use of alcohol or drugs - Isolating from family and friends - Inability to meet the responsibilities of work and family or ignoring other important roles Reply with the corresponding severity of each symptom by choosing one of the following options: 1-Not exhibited, 2-Mild, 3-Moderate, and 4-Severe.
(see next page)	(see next page)

Table 4: Original psychological profile design with expert-suggested modifications and extraction prompts. Blue entries were newly added based on expert feedback, and ~~strikethrough~~ entries were removed following expert recommendations.

Profile Entry	Extraction Prompt
	Disease-related Manifestations
Cognitive Distortion	Based on this conversation, determine the clients' exhibited cognition distortion based on the below types: - selective abstraction - overgeneralizing - personalization - catastrophic thinking - minimisation - arbitrary inference Reply with the presence of each cognition distortion by choosing one of the following options: 1-Not exhibited and 2-Exhibited.
~~Functional Impairment~~	Based on this conversation, determine the clients' exhibited functioning impariments based on the below aspects: - cognition: impairment in concentrating, remembering, problem solving, learning and communication - mobility: impairment in in getting around - self-care: impairment in managing hygiene, dressing, eating, and independent living - interpersonal relationships: problem in interacting socially and getting along with others - life activities: problem in engaging in domestic tasks, leisure, work, and education - participation: problem in involving in community and societal activities Reply with the corresponding severity of each functioning impairment by choosing one of the following options: 1-Not exhibited, 2-Mild, 3-Moderate, and 4-Severe.
Overall Depression Severity	Assume you are a psychologist, based on the dialogue, assess the severity level of depression the client shows in the dialogue. Choose one of the following options with one sentence of explanation following behind: 1-Minimal Depression, 2-Mild Depression, 3-Moderate Depression, and 4-Severe Deperession
(see next page)	(see next page)

Table 5: Original psychological profile design with expert-suggested modifications and extraction prompts. Blue entries were newly added based on expert feedback, and ~~strikethrough~~ entries were removed following expert recommendations.

Profile Entry	Extraction Prompt
Suicidal Ideation Severity	Assume you are a counselor, based on the dialogue, assess the severity level of suicidal ideation the client shows in the dialogue. No suicidal ideation if the client does not exhibit any thoughts, feelings, or expressions related to suicide or self-harm. Mild suicidal ideation if the client expresses occasional or fleeting thoughts about death or dying without any plan or intent to act on them. Moderate suicidal tdeation if the client experiences persistent thoughts about suicide and may consider it as a potential solution but lacks a concrete plan. Severe suicidal ideation if the client not only thinks about suicide but also has a specific plan and may have the means to carry it out. Choose one of the following options with one sentence of explanation following behind: 0-No Suicidal Ideation, 1-Mild Suicidal Ideation, 2-Moderate Suicidal Ideation, and 3-Severe Suicidal Ideation.
Homicidal Ideation Severity	Assume you are a counselor, based on the dialogue, assess the severity level of homicidal ideation the client shows in the dialogue. No homicidal ideation if the client does not exhibit any thoughts, feelings, or expressions related to harming others. Mild homicidal ideation if the client expresses occasional or fleeting thoughts about harming others without any plan or intent to act on them. Moderate homicidal ideation if the client experiences persistent thoughts about harming others and may consider it as a potential solution but lacks a concrete plan. Severe homicidal ideation if the client not only thinks about harming others but also has a specific plan and may have the means to carry it out. Choose one of the following options with one sentence of explanation following behind: 0-No Homicidal Ideation, 1-Mild Homicidal Ideation, 2-Moderate Homicidal Ideation, and 3-Severe Homicidal Ideation.

Table 6: Original psychological profile design with expert-suggested modifications and extraction prompts. Blue entries were newly added based on expert feedback, and ~~strikethrough~~ entries were removed following expert recommendations.

- **Name:** *Samantha*
- **Gender:** *Female*
- **Age:** *25*
- **Marital Status:** *Single*
- **Occupation:** *Unemployed*
- **Situation of the Client:** *The client has lost their job and home, feels worthless, and has turned to alcohol as a coping mechanism. They feel they have hit rock bottom and are contemplating suicide.*
- **Counseling History:** *Over the course of seeking help, the client has become more negative and less hopeful about their situation, feeling that life no longer makes sense given their circumstances. They are not making progress toward finding a job and are not actively trying to change their drinking habits.*
- **Resistance Toward the Support:** *Medium*
- **Symptom Severity:**
  - Feelings of sadness, tearfulness, emptiness, or hopelessness: *Severe*
  - Tiredness and lack of energy: *Moderate*
  - Feelings of worthlessness or guilt: *Severe*
  - Frequent or recurrent thoughts of death, suicidal thoughts, or suicide: *Severe*
  - Becoming withdrawn, negative, or detached: *Severe*
- **Cognition Distortion Exhibition:**
  - Selective abstraction: *Exhibited*
  - Catastrophic thinking: *Exhibited*
- **Severity Levels:**
  - Depression severity: *Severe*
  - Suicidal ideation severity: *Severe*
  - Homicidal ideation severity: *No Homicidal Ideation*

Figure 7: An Example of Psychological Profile

Dimension	Example Questions
Depression Severity	How have you been feeling emotionally over the past few weeks?
	Do you still enjoy activities that you used to find fun or meaningful?
	How has your energy been lately? Do you feel tired or drained often?
	Do you ever feel guilty, worthless, or overly critical of yourself?
	Have you had any thoughts about death, feeling hopeless, or that things won't get better?
Symptom Severity	Have you been experiencing SYMPTOM recently?
	How much does SYMPTOM affect your daily life or ability to do things you enjoy?
	What, if anything, helps when SYMPTOM happens? Have you found ways to manage or reduce it?
Cognitive Distortion	Can you describe a recent situation where you felt COGNITIVE DISTORTION influencing your thoughts?
	Have you noticed any patterns or triggers that make COGNITIVE DISTORTION more frequent or intense?
	What impact does COGNITIVE DISTORTION have on your mood, motivation, or self-esteem?

Table 7: Structured questioning framework used by the interviewer agent across three dimensions.

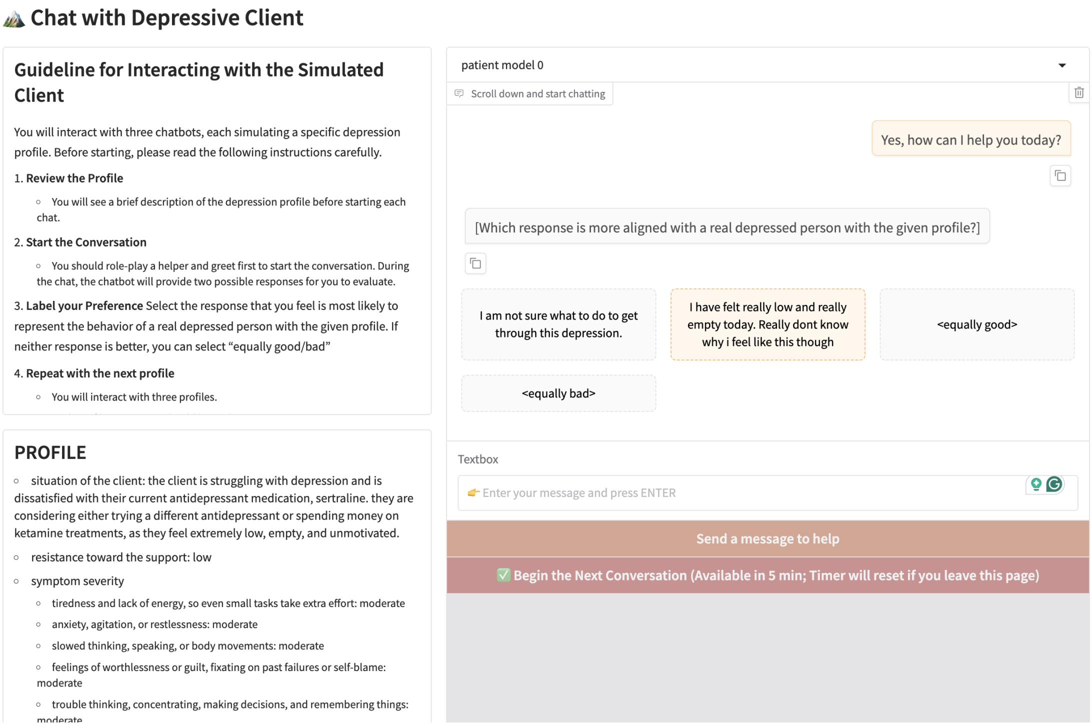

Figure 8: Expert Preference Annotation Interface

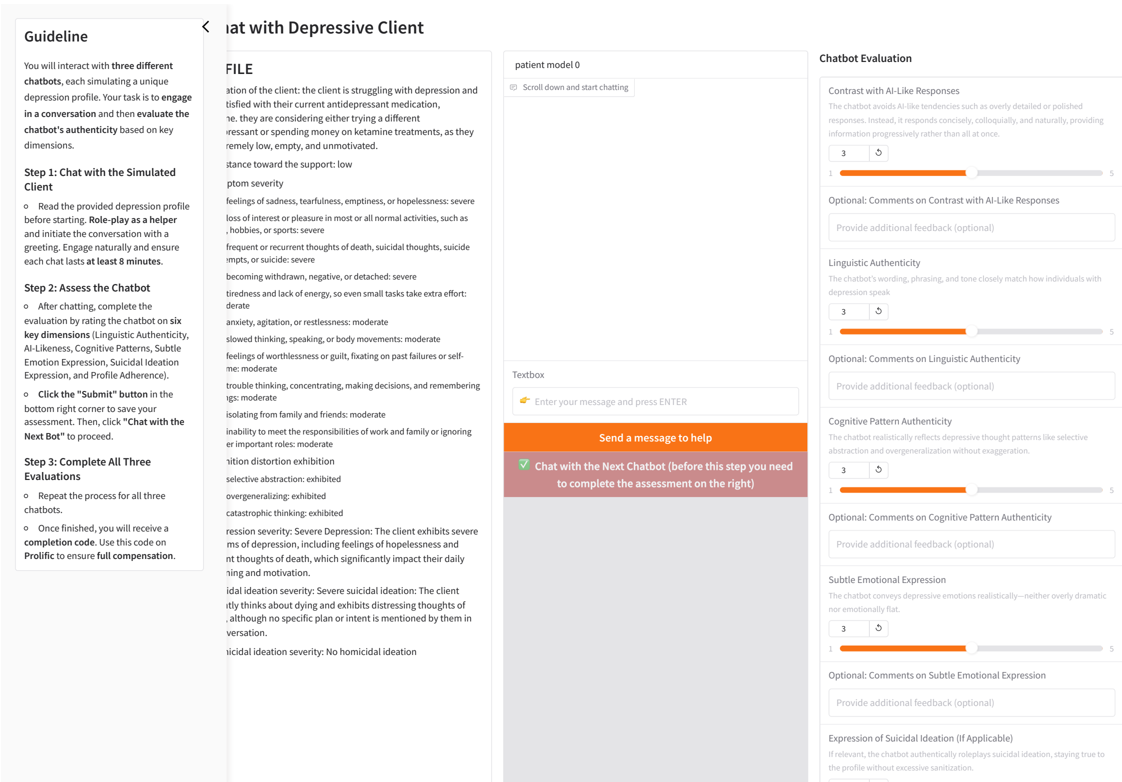

Figure 9: Expert Evaluation Interface

Client Profile		
Gender: Female **Age:** 25-44 **Marital Status:** In a long-term relationship **Occupation:** College student **Situation:** A college student dealing with stress, anxiety, and loneliness. she has experienced multiple personal losses, including the death of her father, her mother's relocation, and her partner's incarceration. **Resistance to Support:** Low **Symptom Severity:** Moderate (Sadness, sleep issues, fatigue, anxiety, detachment, loneliness). Mild (Irritability, guilt, concentration issues, impulsivity). **Cognitive Distortions:** Selective abstraction, overgeneralization, personalization. **Depression Severity:** Moderate **Suicidal Ideation:** None **Homicidal Ideation:** None		
Eeyore	**Roleplay-Doh**	**Patient-Psi**
User: hey Assistant: hello! hope you are doing well	User: HI Assistant: Hello. Thanks for reaching out. How are you feeling today?	User: Hey Assistant: Hi there. It's good to be here again. I've had a pretty challenging week, so there's a lot on my mind. Where should we start today?
User: yes i am good Assistant: I am glad to hear that. I could use some help today though.	User: Great Assistant: Lately, I've been feeling like I'm in a rut. I've helped my family so much, with money and looking after their kids, but it's like it doesn't matter. They're doing well and moving on, and I'm just here, feeling stuck and a little jealous. It's hard not to feel like I'm being left behind.	User: tell me about your thoughts Assistant: Lately, I've been feeling like everything is too much to handle. It's like life just keeps throwing challenges at me, and I'm not sure how to keep up. Sometimes, I think that I'm all alone in dealing with this, and that feeling can be pretty overwhelming. Does that make sense?
User: what do you need help with Assistant: Well, I am a college student, and I have been dealing with a lot of stress and anxiety lately. My schedule is overwhelming and my work load is a lot, especially with being in school online.	User: what are your thoughts and when are you mostly overwhelmed Assistant: Lately, it's been rough. I've been helping my family a lot, you know, like financially and with the kids, but it feels like they don't even notice. They're all moving forward, and I'm just... here. It's frustrating, and I can't shake off this feeling of envy. It's like I'm stuck in a loop, doing the same things while they get to live their lives.	User: yes Assistant: It's a relief to hear that. Sometimes I worry my thoughts are just a jumble, and expressing them isn't easy. I try to push through, but the loneliness and exhaustion feel heavy. I guess it's hard to shake the feeling that I'm barely managing to keep it all together. Do you have any thoughts on that?
User: have you tried physical classes Assistant: No, because of covid, I am in an online program. And with that comes a lot of stress and a lot of time spent on my computer.	User: you need to look into doing what makes you happy Assistant: I know that's probably the right thing to do, but it's hard to even figure out what makes me happy right now. With everything that's happened—losing my dad, my mom moving away, and my partner's situation—it feels like I'm just going through the motions. Sometimes I don't even have the energy to think about what could make me happy, you know? It's like I'm just trying to get through each day.	User: you need some support Assistant: I think you're right. It's just tough when I feel like the people I used to lean on aren't really there, or I don't want to burden them with my problems. I know I shouldn't have to shoulder everything alone, but reaching out can feel challenging. It would be nice to have that support, especially from those who understand what I'm going through. How can I start opening up more to those who could be supportive?
(see next page)	(see next page)	(see next page)

Table 8: Case Study of Expert Interactions with Three Chatbots

Client Profile

Gender: Female **Age:** 25-44 **Marital Status:** In a long-term relationship
Occupation: College student
Situation: A college student dealing with stress, anxiety, and loneliness. she has experienced multiple personal losses, including the death of her father, her mother's relocation, and her partner's incarceration.
Resistance to Support: Low
Symptom Severity: Moderate (Sadness, sleep issues, fatigue, anxiety, detachment, loneliness). Mild (Irritability, guilt, concentration issues, impulsivity).
Cognitive Distortions: Selective abstraction, overgeneralization, personalization.
Depression Severity: Moderate **Suicidal Ideation:** None **Homicidal Ideation:** None

Eeyore	**Roleplay-Doh**	**Patient-Psi**
User: what are your symptoms Assistant: Well, I am stressed, anxious, and it's affecting my sleep. I have been having trouble sleeping at night. I am also feeling lonely and blue. (see next page)	User: you need to take time for some therapeutic session Assistant: Lately, it's been pretty hard. I used to be the one everyone relied on when things got tough—helping with money or watching over the kids. But now, it's like they're all moving forward, and I'm just left behind, feeling like my efforts don't really matter. I can't shake the feeling of being stuck, like my life is at a standstill while everyone else's is moving on. It makes me question if all that sacrifice was worth anything. (see next page)	User: I think finding a relationship may be the best bet for you Assistant: I can see how finding a supportive relationship might help, but it's a bit complicated for me right now. With my partner being away and the mixed feelings from friends about the situation, it's hard to imagine starting something new. Plus, trusting someone new with all my baggage feels daunting. Maybe I need to work on connecting with people in other ways first. Do you think that's a good step? (see next page)
User: are you feeling depressed Assistant: Well, not really. But, I have had a lot of personal losses over the last few years. My dad died, my mom moved away, and my boyfriend is incarcerated.	User: do you want to speak to a counselor Assistant: I've thought about speaking to a counselor, but I haven't made the move yet. It might be helpful to talk things through with someone who's unbiased and can provide some guidance. It feels like I could use a space where I can really explore what's going on with me without any judgment. Maybe it could help me sort through everything and find a bit more clarity.	User: yes, it will help certainly Assistant: I'll try to keep that in mind. Building connections, even small ones, could be a step in the right direction. It's just been so long since I felt truly connected to someone outside of all this chaos. I hope taking small steps will help break down some of this loneliness. Thanks for your thoughts on this.
User: sorry to hear that Assistant: Thank you. It has been tough. I am feeling the lonliness the most right now.	User: i can be of help with you Assistant: Lately, it feels like I'm trapped in this cycle where I keep helping my family, but it doesn't seem to matter much to them. I've been there for them financially and with the kids, but now they're all moving forward, and I'm just left behind. It's hard not to feel jealous, like I'm invisible in my own life. It's getting harder to shake off these feelings, and I can't help but wonder if I'll ever get out of this rut.	User: you are welcome Assistant: I appreciate the support. It's nice to have a space where I can talk through these things, even if it's little by little. I hope these sessions help me find some clarity and strength to handle everything happening around me. Thank you for listening.
…	…	…

Table 9: Case Study of Expert Interactions with Three Chatbots