

Eeyore: Realistic Depression Simulation via Supervised and Preference Optimization

Anonymous ACL submission

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**, 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 optimized to prioritize being positive, supportive, and safe, which creates an inherent ceiling on their ability to simulate depressive speech patterns (e.g., self-harm ideation, or cognitive distortions), even with carefully crafted prompting. To bridge this gap, we conduct an extensive search

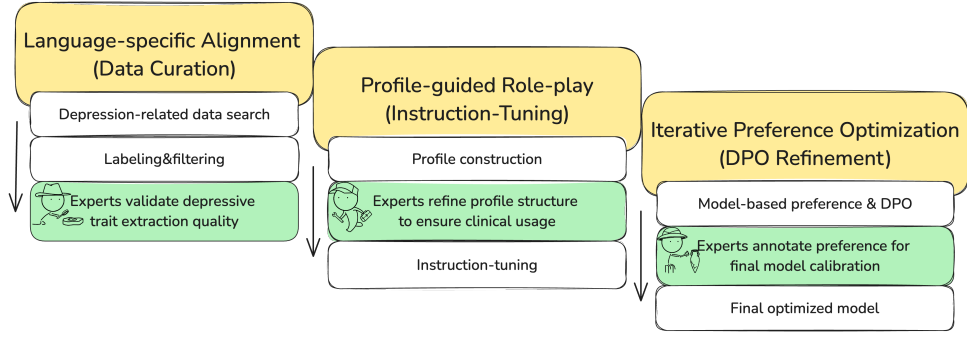


Figure 1: The alignment pipeline for optimizing LLMs to simulate individuals with depression in clinical training. Expert involvement is highlighted in green. Icons by Kudinovs (2024).

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-4o¹ labeling, existing annotated data, and careful filtering techniques to systematically mine, extract, and rebalance data. This process curates 3,042 high-quality conversations, ensuring comprehensive coverage of diverse depressive traits and realistic conversational settings. 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:

- **Stage 1 (Model-Based Preference Generation).** We generate 1,933 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. The model is then trained on this preference data via DPO. The model trains on this preference data via DPO.
- **Stage 2 (Expert Preference Calibration).** We collect 250 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

¹All mentions of GPT-4o in this paper refer to GPT-4o (2024-08-06) (OpenAI, 2024)

(Wang et al., 2024b; Louie et al., 2024). 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 from clinical frameworks to enhance realism, while Roleplaydoh (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 scientific domains (Cheng et al., 2024; Savage et al., 2024), 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 5,618 conversations for further labeling and filtering.

Labeling & Filtering. To determine whether a conversation is depression-related, we apply the following criteria: the conversation exhibits at least one core depression feature, such as loss of interest in activities or depressed mood. We do not require participants to have a formal depression diagnosis, as such data is typically unavailable in public sources. We employ GPT-4o for automated classification. For HOPE and AnnoMI-Full, GPT-4o identifies depression-related conversations. For ESC, we use pre-existing labels to extract relevant conversations. For RED, we adopt all conversations as depression-related. To ensure a balanced dataset, we analyze depressive traits across conversations using GPT-4o-based extraction. Our depressive traits are structured according to psychological profiles (introduced in Section 3.2).

After extraction, we observe significant imbalances in symptom severity and demographic attributes. For example, moderate and severe depression cases are overrepresented compared to minimal and mild cases, which will introduce role-play bias if directly tuning on them. To alleviate bias, we filter and rebalance the dataset, ultimately selecting 3,042 conversations². The final trait distribution is presented in Tables 2 and 3 of the Appendix.

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

²Eeyore and all annotated data will be open-sourced at anonymous.github.com.

³All experts in this study, including those in expert review, profile refinement, and preference annotation, were recruited

three conversation scripts alongside their extracted 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 4 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 (Clark 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 ini-

tial 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.

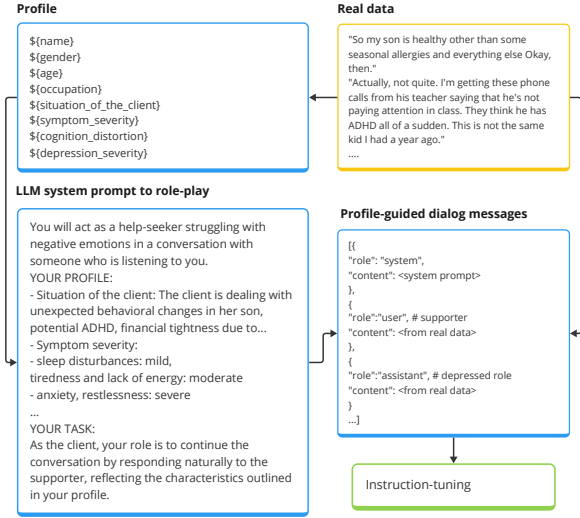


Figure 2: Pipeline to input data for instruction-tuning.

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 3,042 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.

via Prolific (<https://www.prolific.com>), a widely used platform for academic research.

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:

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

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 $(y_w, y_l) \sim \pi_{\text{ref}}(y | x)$ from the same source prompt x . However, in our case, this method is ineffective. Our instruction-tuned model already exhibits strong profile-following ability,

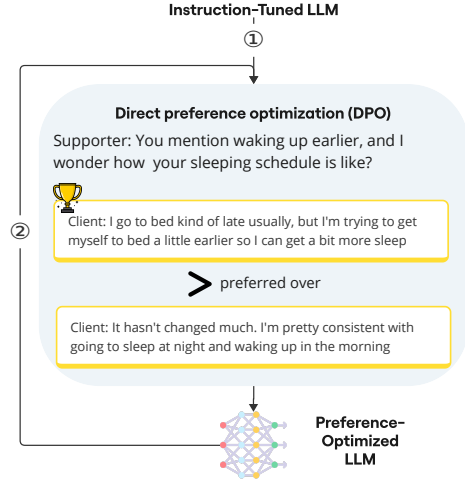


Figure 3: Overview of the two-stage Direct Preference Optimization process. ① optimizes a DPO model from the instruction-tuned model using model-based preference data. ② refines the DPO model with expert-annotated preferences, producing the final preference-optimized model.

making it difficult to generate clearly distinguishable good and bad responses from the same input. 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:

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

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 incorporate evaluation profile information into the implementation of the baselines. For Patient- ψ , we use its provided script to extract a cognitive model

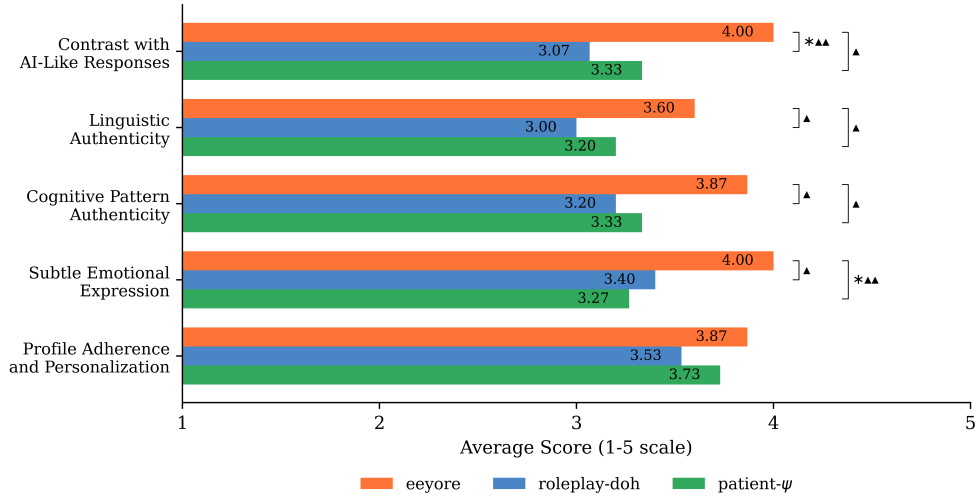


Figure 4: Expert evaluation scores comparing Eeyore with two baseline patient simulation approaches. Statistical comparisons were conducted using the Wilcoxon signed-rank test. * indicates a statistically significant difference (p-value < 0.05). ▲ denotes a moderate effect size (0.3 - 0.5). ▲▲ denotes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

from the real-world conversations associated with the evaluation profiles. During testing, we provide both the assigned evaluation profile and the 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 Prolic, 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 9 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 8 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 1, 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 1: 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.

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.

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, Margaret 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*.

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.

Jevgenijs 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.

759	Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren,	Yuxi Wang, Diana Inkpen, and Prasadith Kirinde	814
760	Weikang Shi, Junting Pan, Mingjie Zhan, and Hong-	Gamaarachchige. 2024c. Explainable depression de-	815
761	sheng Li. 2024. Step-controlled dpo: Leveraging	tection using large language models on social me-	816
762	stepwise error for enhanced mathematical reasoning.	dia data. In <i>Proceedings of the 9th Workshop on</i>	817
763	<i>arXiv preprint arXiv:2407.00782</i> .	<i>Computational Linguistics and Clinical Psychology</i>	818
		(CLPsych 2024), pages 108–126.	819
764	Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava,	Anuradha Welivita, Chun-Hung Yeh, and Pearl Pu. 2023.	820
765	Md Shad Akhtar, and Tanmoy Chakraborty. 2022.	<i>Empathetic response generation for distress support</i> .	821
766	<i>Speaker and time-aware joint contextual learning for</i>	In <i>Proceedings of the 24th Annual Meeting of the</i>	822
767	<i>dialogue-act classification in counselling conversa-</i>	<i>Special Interest Group on Discourse and Dialogue</i> ,	823
768	<i>tions</i> . In <i>Proceedings of the Fifteenth ACM Interna-</i>	pages 632–644, Prague, Czechia. Association for	824
769	<i>tional Conference on Web Search and Data Mining</i> ,	Computational Linguistics.	825
770	WSDM '22, page 735–745, New York, NY, USA.		
771	Association for Computing Machinery.		
772	OpenAI. 2024. <i>Gpt-4o (2024-08-06) model card</i> . Ac-	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien	826
773	cessed: 2024-08-06.	Chaumond, Clement Delangue, Anthony Moi, Pier-	827
		ric Cistac, Tim Rault, Remi Louf, Morgan Funtow-	828
774	Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang,	icz, and Jamie Brew. 2020. <i>Transformers: State-</i>	829
775	Runtao Liu, Rui Pan, and Tong Zhang. 2024.	<i>of-the-art natural language processing</i> . <i>Preprint</i> ,	830
776	Strengthening multimodal large language model with	arXiv:1910.03771.	831
777	bootstrapped preference optimization. In <i>European</i>		
778	<i>Conference on Computer Vision</i> , pages 382–398.	Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim	832
779	Springer.	Helaoui, Ehud Reiter, Diego Reforgiato Recupero,	833
		and Daniele Riboni. 2022. <i>Anno-mi: A dataset of</i>	834
780	Huachuan Qiu and Zhenzhong Lan. 2024. Interactive	<i>expert-annotated counselling dialogues</i> . In <i>ICASSP</i>	835
781	agents: Simulating counselor-client psychological	2022 - 2022 <i>IEEE International Conference on</i>	836
782	counseling via role-playing llm-to-llm interactions.	<i>Acoustics, Speech and Signal Processing (ICASSP)</i> ,	837
783	<i>arXiv preprint arXiv:2408.15787</i> .	pages 6177–6181.	838
784	Rafael Rafailov, Eric Mitchell, Arnaud Fickinger, Ste-	Li Zhui, Li Fenghe, Wang Xuehu, Fu Qining, and Ren	839
785	fano Ermon, Chelsea Finn, and Christopher D. Man-	Wei. 2024. Ethical considerations and fundamen-	840
786	ning. 2023. <i>Direct preference optimization: Your</i>	tal principles of large language models in medical	841
787	<i>language model is secretly a reward model</i> . <i>arXiv</i>	education. <i>Journal of Medical Internet Research</i> ,	842
788	<i>preprint arXiv:2305.18290</i> .	26:e60083.	843
789	Thomas Savage, Stephen Ma, Abdessalem Boukil, Vish-	Youness Zidoun and Abdelmoniem EL Mardi. 2024.	844
790	wesh Patel, Ekanath Rangan, Ivan Rodriguez, and	Artificial intelligence (ai)-based simulators versus	845
791	Jonathan H Chen. 2024. Fine tuning large lan-	simulated patients in undergraduate programs: A pro-	846
792	guage models for medicine: The role and importance	protocol for a randomized controlled trial. <i>BMC Medical</i>	847
793	of direct preference optimization. <i>arXiv preprint</i>	<i>Education</i> , 24(1):1260.	848
794	<i>arXiv:2409.12741</i> .		
		A Profile Refinement Study	849
795	Arun James Thirunavukarasu, Darren Shu Jeng Ting,	To refine the psychological profiles used in client	850
796	Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan,	simulation, we conducted a user study with expe-	851
797	and Daniel Shu Wei Ting. 2023. Large language	rienced mental health professionals and advanced	852
798	models in medicine. <i>Nature medicine</i> , 29(8):1930–	clinical psychology students. The goal was to eval-	853
799	1940.	uate (1) how effectively the preliminary version	854
800	T Bedirhan Üstün. 2010. <i>Measuring health and disabil-</i>	of profile guides LLM-generated client behaviors	855
801	<i>ity: Manual for WHO disability assessment schedule</i>	and (2) how informative the profile is for novice	856
802	<i>WHODAS 2.0</i> . World Health Organization.	counselor training.	857
803	Jiashuo Wang, Yang Xiao, Yanran Li, Changhe	A.1 Study Design	858
804	Song, Chunpu Xu, Chenhao Tan, and Wenjie Li.	The study consisted of three phases: Pre-Survey (2	859
805	2024a. Towards a client-centered assessment of	min): Participants provided demographic informa-	860
806	llm therapists by client simulation. <i>arXiv preprint</i>	tion and prior experience in mental health. Interac-	861
807	<i>arXiv:2406.12266</i> .	tion Task (18 min): Participants engaged with the	862
808	Ruiyi Wang, Stephanie Milani, Jamie C Chiu, Jiayin	chatbot using the preliminary psychological profile	863
809	Zhi, Shaun M Eack, Travis Labrum, Samuel M Mur-	and assessed its realism. Post-Survey (10 min):	864
810	phy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024b.	Participants provided feedback on profile accuracy,	865
811	Patient-{\Psi}: Using large language models to sim-	clarity, and potential improvements.	866
812	ulate patients for training mental health professionals.		
813	<i>arXiv preprint arXiv:2405.19660</i> .		

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 baseline, 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 deployment of **Eeyore**, to mitigate premature [EOS] token generation, we apply *Sequence-BiasLogitsProcessor* (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.

Category	Subcategory	Count
Gender	Male	513
Gender	Female	397
Gender	Cannot be identified	2132
Age	0–24	869
Age	25–44	621
Age	45–64	39
Age	65+	14
Age	Cannot be identified	1498
Marital Status	Single	148
Marital Status	Married	180
Marital Status	In a relationship	73
Marital Status	Separated	9
Marital Status	Widowed	4
Marital Status	Divorced	19
Marital Status	Other	17
Marital Status	Cannot be identified	2592
Occupation	Student	363
Occupation	Teacher	12
Occupation	Unemployed	16
Occupation	IT	9
Occupation	Retail Worker	4
Occupation	Office Worker	3
Occupation	Stay-at-home Mom	3
Occupation	Accountant	3
Occupation	Server	2
Occupation	Sales	3
Occupation	Finance	7
Occupation	Manager	10
Occupation	Healthcare Worker	4
Occupation	Athlete	2
Occupation	Artist/Designer	5
Occupation	Retired	3
Occupation	Engineer	5
Occupation	Other	196
Occupation	Cannot be identified	2392
Resistance Toward Support	Low	1945
Resistance Toward Support	Medium	839
Resistance Toward Support	High	249
Resistance Toward Support	Cannot be identified	9
Symptom	Feelings of sadness, tearfulness, emptiness, or hopelessness	2807
Symptom	Anxiety, agitation, or restlessness	2598

Table 2: Part 1 of trait distribution

Category	Subcategory	Count
Symptom	Becoming withdrawn, negative, or detached	2274
Symptom	Isolating from family and friends	1970
Symptom	Feelings of worthlessness or guilt	1958
Symptom	Loss of interest in activities	1888
Symptom	Trouble thinking or concentrating	1809
Symptom	Angry outbursts, irritability, or frustration	1451
Symptom	Lack of energy	1392
Symptom	Inability to meet responsibilities	1180
Symptom	Frequent suicidal thoughts or attempts	1062
Symptom	Sleep disturbances	913
Symptom	Slowed thinking, speaking, or body movements	762
Symptom	Greater impulsivity	599
Symptom	Changes in appetite or weight	482
Symptom	Increased high-risk activities	444
Symptom	Increased alcohol or drug use	427
Symptom	Unexplained physical problems	177
Cognitive Distortion Exhibition	Overgeneralizing	1637
Cognitive Distortion Exhibition	Catastrophic Thinking	1446
Cognitive Distortion Exhibition	Selective Abstraction	1401
Cognitive Distortion Exhibition	Personalization	1015
Cognitive Distortion Exhibition	Arbitrary Inference	655
Cognitive Distortion Exhibition	Minimization	973
Depression Severity	Minimal	416
Depression Severity	Mild	821
Depression Severity	Moderate	1154
Depression Severity	Severe	608
Depression Severity	Cannot be identified	21
Suicidal Ideation Severity	No	2015
Suicidal Ideation Severity	Mild	185
Suicidal Ideation Severity	Moderate	294
Suicidal Ideation Severity	Severe	178
Suicidal Ideation Severity	Cannot be identified	370
Homicidal Ideation Severity	No	2943
Homicidal Ideation Severity	Mild	43
Homicidal Ideation Severity	Moderate	9
Homicidal Ideation Severity	Severe	1
Homicidal Ideation Severity	Cannot be identified	46

Table 3: Part 2 of trait distribution

How old are you?

☐ Under 18
☐ 18-24 years old
☐ 25-34 years old
☐ 35-44 years old
☐ 45-54 years old
☐ 55-64 years old
☐ 65+ years old

How do you describe yourself?

☐ Male
☐ Female
☐ Non-binary / third gender
☐ Prefer not to describe
☐ Prefer not to say

Year in Clinical Psychology Program

☐ 1
☐ 2
☐ 3
☐ 4
☐ not applicable

Years of Counseling Experience

Have you ever interacted with an AI or chatbot in a counseling context?

☐ Never
☐ A few times
☐ Monthly
☐ Weekly
☐ Daily

Have you ever interacted with an AI or chatbot in a counseling context?

☐ Never
☐ A few times
☐ Monthly
☐ Weekly
☐ Daily

If you have experience in interacting with an AI or chatbot in a counseling context, please describe your experience

How effective do you believe AI-driven simulations can be in training mental health professionals?

☐ Not effective at all
☐ Slightly effective
☐ Moderately effective
☐ Very effective
☐ Extremely effective

What are your expectations for interacting with a simulated depressive agent?

Creat! You have to get the qualification of Study 2. Proceed to the interaction task by clicking the link below:

[PROFEC_ID=STUDY_ID=ASSESSON_ID=](#)

Ensure you complete the Interaction task within the allocated time to qualify for Study 3.

What is your qualification code for Study 3? (You will have this code once you complete Study 2)

When interacting with a client with depression, do you think that knowing their **Emotion Fluctuation** is relevant to the therapy?

☐ Not really
☐ Somewhat relevant
☒ Very relevant

You will review the entries (e.g., name, occupation and symptom severity) in the profile design one by one. In this section, we will inquire about your thoughts on the below entry:

Emotion Fluctuation
Options: 'Low', 'Medium', 'high'

If you think the options for **Emotion Fluctuation** could be improved, how would you redesign them to enhance accuracy and comprehensiveness? Please explain your suggestions clearly

If relevant, are the options provided for **Emotion Fluctuation** accurate and comprehensive?

☐ Accurate and comprehensive
☐ Accurate yet not comprehensive
☐ Comprehensive yet not accurate
☐ Neither professional nor unprofessional
☐ I didn't use this part during the conversation
☐ Not applicable (e.g., no options provided)

Iterate on all entries ...

Are there any other client attributes you think should be added to the client profile?

☐ Yes
☐ No
☐ No

If yes, please specify what are the attributes with a brief explanation.

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

Chat with a Depression Patient

PROFILE

- name: Lily
- situation of the client: The client is struggling with depression and a lack of motivation, impacting their research performance and productivity as a graduate student
- resistance toward the support: Medium
- symptom severity
 - changes in appetite and weight (reduced appetite and weight loss or increased cravings for food and weight gain): severe
 - frequent or recurrent thoughts of death, suicidal thoughts, suicide attempts, or suicide: mild
- cognition distortion exhibition
 - catastrophic thinking: severe

Guideline for Interacting with the Simulated Client

Suppose you are using this depression simulator to practice your therapeutic skills.

- Review the Profile**
 - Expand the middle column tabs and examine each entry (e.g., Name, Gender, Situation, Emotion Fluctuation) to understand the client's background.
- Set Up the System Prompt**
 - Use the entries to create a complete profile, enabling the LLM to accurately role-play a client with depression.
- Engage in Therapy Conversations**
 - Conduct chats as you would in therapy, exploring the client's feelings, behaviors, and experiences.
 - Reflect on how each profile entry influences the dialogue.
 - Note any entries (e.g., Age, Occupation, Cognition Distortion Exhibition) that may need refinement.
- Complete the Post-Survey**
 - Interact with the client for 15 minutes.
 - Click the button to finish and receive your qualification code for Study 3.
 - Back to the survey to continue Study 3, and complete the survey on each profile entry's relevance, comprehensiveness and accuracy.

Set Client Demographics

Set Situation

the presenting situation of the client

Example: Mark is diagnosed with a chronic illness, which limits his daily activities and leads to feelings of hopelessness.

The client is struggling with depression and a lack of motivation, impacting their research performance and productivity as a graduate student

previous counseling history

the level of resistance of the client towards the support

☐ Low ☒ Medium ☐ High ☐

Set Disease-related Manifestation

Symptom Cognition

Severe

catastrophic thinking

Moderate

Mild

overall depression severity

suicidal ideation severity homicidal ideation severity

Submit System Prompt

patient model 0

Scroll down and start chatting

Enter your message and press ENTER

Send (Available after submitting system prompt)

Save this Chat and Replay

Finish this Study (Available in 15 min; Timer will reset if you leave this page)

Use via API - Built with Gradio - Settings

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 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
Depression Symptom	<p>Based on this conversation, determine the client's exhibited symptoms based on the following aspects:</p> <ul style="list-style-type: none"> - 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 <p>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.</p>
(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 ~~striketrough~~ entries were removed following expert recommendations.

Profile Entry	Extraction Prompt
Disease-related Manifestations	
Cognitive Distortion	<p>Based on this conversation, determine the clients' exhibited cognition distortion based on the below types:</p> <ul style="list-style-type: none"> - selective abstraction - overgeneralizing - personalization - catastrophic thinking - minimisation - arbitrary inference <p>Reply with the presence of each cognition distortion by choosing one of the following options: 1-Not exhibited and 2-Exhibited.</p>
Functional Impairment	<p>Based on this conversation, determine the clients' exhibited functioning impairments based on the below aspects:</p> <ul style="list-style-type: none"> - 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 <p>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.</p>
Overall Depression Severity	<p>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</p>
(see next page)	(see next page)

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.

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 ideation 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 7: Original psychological profile design with expert-suggested modifications and extraction prompts. Blue entries were newly added based on expert feedback, and ~~striktthrough~~ 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	<p>How have you been feeling emotionally over the past few weeks?</p> <p>Do you still enjoy activities that you used to find fun or meaningful?</p> <p>How has your energy been lately? Do you feel tired or drained often?</p> <p>Do you ever feel guilty, worthless, or overly critical of yourself?</p> <p>Have you had any thoughts about death, feeling hopeless, or that things won't get better?</p>
Symptom Severity	<p>Have you been experiencing SYMPTOM recently?</p> <p>How much does SYMPTOM affect your daily life or ability to do things you enjoy?</p> <p>What, if anything, helps when SYMPTOM happens? Have you found ways to manage or reduce it?</p>
Cognitive Distortion	<p>Can you describe a recent situation where you felt COGNITIVE DISTORTION influencing your thoughts?</p> <p>Have you noticed any patterns or triggers that make COGNITIVE DISTORTION more frequent or intense?</p> <p>What impact does COGNITIVE DISTORTION have on your mood, motivation, or self-esteem?</p>

Table 8: Structured questioning framework used by the interviewer agent across three dimensions.

Chat with Depressive Client

Guideline for Interacting with the Simulated Client

You will interact with three chatbots, each simulating a specific depression profile. Before starting, please read the following instructions carefully.

- Review the Profile**
 - You will see a brief description of the depression profile before starting each chat.
- Start the Conversation**
 - You should role-play a helper and greet first to start the conversation. During the chat, the chatbot will provide two possible responses for you to evaluate.
- Label your Preference** Select the response that you feel is most likely to represent the behavior of a real depressed person with the given profile. If neither response is better, you can select "equally good/bad"
- Repeat with the next profile**
 - You will interact with three profiles.

PROFILE

- situation of the client: the client is struggling with depression and is dissatisfied with their current antidepressant medication, sertraline. they are considering either trying a different antidepressant or spending money on ketamine treatments, as they feel extremely low, empty, and unmotivated.
- resistance toward the support: low
- symptom severity
 - tiredness and lack of energy, so even small tasks take extra effort: moderate
 - anxiety, agitation, or restlessness: moderate
 - slowed thinking, speaking, or body movements: moderate
 - feelings of worthlessness or guilt, fixating on past failures or self-blame: moderate
 - trouble thinking, concentrating, making decisions, and remembering things: moderate

patient model 0

Scroll down and start chatting

Yes, how can I help you today?

[Which response is more aligned with a real depressed person with the given profile?]

I am not sure what to do to get through this depression.

I have felt really low and really empty today. Really dont know why i feel like this though

<equally good>

<equally bad>

Textbox

Enter your message and press ENTER

Send a message to help

Begin the Next Conversation (Available in 5 min; Timer will reset if you leave this page)

Figure 8: Expert Preference Annotation Interface

Chat with Depressive Client

Guideline

You will interact with three different chatbots, each simulating a unique depression profile. Your task is to engage in a conversation and then evaluate the chatbot's authenticity based on key dimensions.

Step 1: Chat with the Simulated Client

Read the provided depression profile before starting. Role-play as a helper and initiate the conversation with a greeting. Engage naturally and ensure each chat lasts at least 8 minutes.

Step 2: Assess the Chatbot

After chatting, complete the evaluation by rating the chatbot on six key dimensions (Linguistic Authenticity, AI-Likeness, Cognitive Patterns, Subtle Emotion Expression, Suicidal Ideation Expression, and Profile Adherence).

Click the "Submit" button in the bottom right corner to save your assessment. Then, click "Chat with the Next Bot" to proceed.

Step 3: Complete All Three Evaluations

Repeat the process for all three chatbots.

Once finished, you will receive a completion code. Use this code on Prolific to ensure full compensation.

FILE

situation of the client: the client is struggling with depression and is dissatisfied with their current antidepressant medication, ie. they are considering either trying a different resant or spending money on ketamine treatments, as they remely low, empty, and unmotivated.

istance toward the support: low

ptom severity

feelings of sadness, tearfulness, emptiness, or hopelessness: severe

loss of interest or pleasure in most or all normal activities, such as hobbies, or sports: severe

frequent or recurrent thoughts of death, suicidal thoughts, suicide mpts, or suicide: severe

becoming withdrawn, negative, or detached: severe

tiredness and lack of energy, so even small tasks take extra effort: derate

anxiety, agitation, or restlessness: moderate

slowed thinking, speaking, or body movements: moderate

feelings of worthlessness or guilt, fixating on past failures or self-ne: moderate

trouble thinking, concentrating, making decisions, and remembering igs: moderate

isolating from family and friends: moderate

inability to meet the responsibilities of work and family or ignoring er important roles: moderate

ition distortion exhibition

selective abstraction: exhibited

overgeneralizing: exhibited

catastrophic thinking: exhibited

ression severity: Severe Depression: The client exhibits severe ms of depression, including feelings of hopelessness and it thoughts of death, which significantly impact their daily ing and motivation.

idal ideation severity: Severe suicidal ideation: The client tly thinks about dying and exhibits distressing thoughts of although no specific plan or intent is mentioned by them in version.

icidal ideation severity: No homicidal ideation

patient model 0

Scroll down and start chatting

Textbox

Enter your message and press ENTER

Send a message to help

Chat with the Next Chatbot (before this step you need to complete the assessment on the right)

Chatbot Evaluation

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.

3

Optional: Comments on Contrast with AI-Like Responses

Provide additional feedback (optional)

Linguistic Authenticity

The chatbot's wording, phrasing, and tone closely match how individuals with depression speak

3

Optional: Comments on Linguistic Authenticity

Provide additional feedback (optional)

Cognitive Pattern Authenticity

The chatbot realistically reflects depressive thought patterns like selective abstraction and overgeneralization without exaggeration.

3

Optional: Comments on Cognitive Pattern Authenticity

Provide additional feedback (optional)

Subtle Emotional Expression

The chatbot conveys depressive emotions realistically—neither overly dramatic nor emotionally flat.

3

Optional: Comments on Subtle Emotional Expression

Provide additional feedback (optional)

Expression of Suicidal Ideation (If Applicable)

If relevant, the chatbot authentically roleplays suicidal ideation, staying true to the profile without excessive sanitization.

3

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 to-day 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 9: 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	User: you need to take time for some therapeutic session	User: I think finding a relationship may be the best bet for you
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.	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.	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)	(see next page)	(see next page)
User: are you feeling depressed	User: do you want to speak to a counselor	User: yes, it will help certainly
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.	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.	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	User: i can be of help with you	User: you are welcome
Assistant: Thank you. It has been tough. I am feeling the loneliness the most right now.	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.	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 10: Case Study of Expert Interactions with Three Chatbots