

Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles

Anonymous ACL submission

Abstract

Recent works leverage LLMs to roleplay realistic social scenarios, aiding novices in practicing their social skills. However, simulating sensitive interactions, such as in mental health, is challenging. Privacy concerns restrict data access, and collecting expert feedback, although vital, is laborious. To address this, we develop Roleplay-doh, a novel human-LLM collaboration pipeline that elicits qualitative feedback from a domain-expert, which is transformed into a set of principles, or natural language rules, that govern an LLM-prompted roleplay. A focal the domain of mental health with counselors customizing AI patients as simulated practice partners for novice counselors. After uncovering issues in GPT-4 simulations not adhering to expert-defined principles, we also introduce a novel principle-adherence prompting pipeline which shows 30% improvements in response quality and principle following for the downstream task. Via a user study with 25 counseling experts, we demonstrate that the pipeline makes it easy and effective to create AI patients that more faithfully resemble real patients, as judged by creators and third-party counselors.

1 Introduction

The application of LLMs in simulations holds great potential for a variety of interactive applications, ranging from social skill training systems as AI practice partners (Yang et al., 2024) to prototyping tools that use them as believable proxies of human behavior (Park et al., 2022). However, achieving realistic and reliable simulations remains a significant challenge, due to issues such as caricature (Cheng et al., 2023), bias, and limited domain knowledge. Existing methods for improving LLM simulations such as finetuning (Demasi et al., 2020) can help, but in sensitive application domains like mental health, privacy concerns with obtaining the required data can restrict their feasibility. This

necessitates *experts-in-the-loop* to guide the evaluation and refinement (Chen et al., 2023; Stapleton et al., 2023) of simulations.

However, how to involve experts when improving simulations is an open challenge. Collecting sufficient amounts of binary or preference data from experts for post-training (Christiano et al., 2017; Rafailov et al., 2024) can be tedious and expensive. Experts can guide the prompting of LLM simulations, directly by editing their own prompts or indirectly through testing and think-aloud sessions. However each method has its limitations: domain-experts may not know how to prompt simulations for desired behaviors (Zamfirescu-Pereira et al., 2023); and indirect methods are inefficient as it requires a designer or researcher to translate qualitative insights into prompt-design changes.

As a focal example, consider the problem of creating AI patients that serve as roleplay partners to enable varied and interactive practice opportunities for novice therapists and counselors (Yao et al., 2022). Creating realistic simulations by fine-tuning on mental health data is infeasible because therapy transcripts with real patients is difficult to obtain due to privacy concerns. Naively prompting LLMs fail to resemble typical behaviors of real-patients such as using colloquial language and showing resistance to help, when evaluated by mental health experts (Chen et al., 2023). To date, no system supports counseling experts, who are familiar with real-patient behaviors but are unlikely to have the technical expertise to effectively write prompts, to customize an AI patient themselves.

In response, we aim to enable human-LLM collaboration for realistic simulation by developing a novel interactive tool, called Roleplay-doh, that empowers domain experts to *directly* guide the creation of simulations by providing *qualitative feedback without any explicit prompting*. Our initial tool design adopts an intuitive and effective paradigm for user-driven chatbot assistant de-

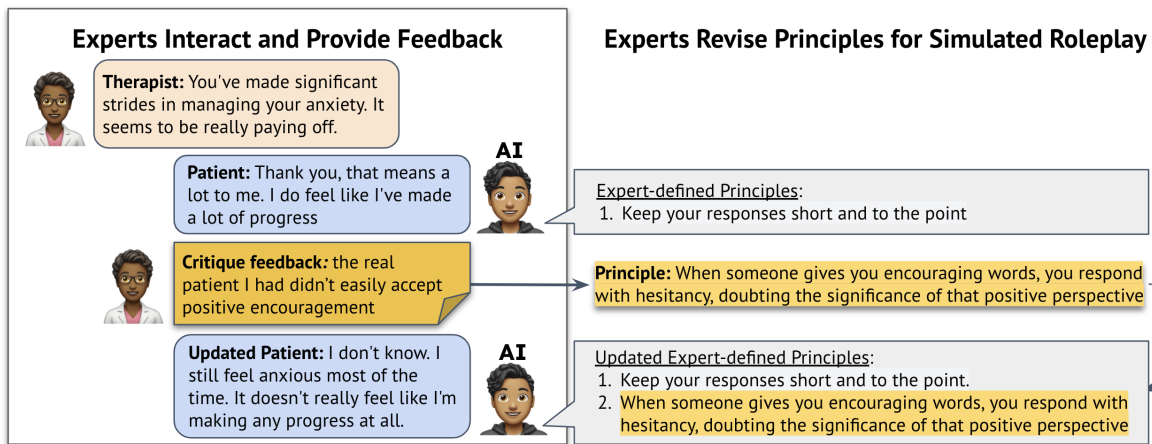


Figure 1: Roleplay-doh empowers an expert counselor to create a customized AI patient intended for other novice counselors to use as a practice partner. While interacting with the AI patient, the expert counselor can provide qualitative feedback which is converted by an LLM into a principle, or a custom rule governing desired roleplay behavior. The principle is appended to the AI Patient's Constitution

sign (Petridis et al., 2023) to help domain-experts drive the design of LLM simulations: experts customize a set of *principles*, or rules written in natural language that govern its behavior (Bai et al., 2022)—by (1) interactively critiquing responses in natural language that then (2) gets transformed by an LLM into well-formulated principles describing how the LLM simulation should act from now on for example, "Respond to encouraging words with hesitation, doubting their significance" (Fig 1). The principles are then used along with a persona description to generate roleplay responses.

In our initial tests of the tool with expert-counselors, we found that even with expert refinement via principles, the LLM- simulations had difficulty delivering high-quality responses consistently. Our analysis of GPT-4 prompted simulation revealed that in 20% of responses, the simulation had difficulty adhering to multipart principles and misapplying those principles that are only applicable in specific contexts e.g., *only when the therapist provides encouraging words*. To resolve these issues, we introduce a novel **principle-adherence pipeline** in the final tool design. The first stage in the pipeline decomposes multipart and contextual principles into a set of yes/no questions that are easier to judge, and the second stage assesses the applicability of each simplified principle to the current scenario before self-refining (Madaan et al., 2023) the AI patient response as required.

We conducted a detailed evaluation of Roleplay-doh to assess its human-LLM collaboration pipeline, focusing on how expert feedback helps

develop more authentic AI patients for training. In a within-subjects study involving 25 expert counselors, participants created AI patients either by describing real-patient scenarios or by using Roleplay-doh to refine simulation principles. The results show that Roleplay-doh enables counselors to produce AI patients that are more authentic, closely resemble real cases, and are better prepared for training use, as judged by creators and third-party counselors. Further, our principle-adherence pipeline achieves the highest principle following (W: 35%; L: 5%) and dialogue consistency (W: 35%; L: 10%) compared to all ablations, where preferences are made against a baseline that does not self-refine its output. This work highlights the limitations of existing LLM simulation systems in specialized, data-scarce domains like mental health counseling, and designs and validates a tool that enables expert counselors to directly customize LLM simulations of AI patients. Since Roleplay-doh does not contain any components specifically tailored for the domain of mental health, we hypothesize that the tool can be used to build realistic LLM simulations for a wide variety of domains, with appropriate expert feedback.

2 Related Work

Utility of Simulated Partners Simulated partners are used to give social skill learners the needed practice opportunities that textbook knowledge cannot provide. Past education software develops digital patient simulations to make simulated partners more accessible (Othlinghaus-Wulhorst and Hoppe, 2020) but their tailored dialogue trees limit the con-

149 texts for practice. LLMs can overcome this issue
150 by being flexibly configured to convincingly sim-
151 ulate a diverse set of personas (Park et al., 2022)
152 and characters (Park et al., 2023) and generate re-
153 sponses in a range of contexts. Researchers have
154 thus explored their application for simulation train-
155 ing for teaching (Markel et al., 2023), conflict res-
156 olution (Shaikh et al., 2023), and counseling (De-
157 masi et al., 2020; Tanana et al., 2019). Previous
158 work has proposed methods to simulate diverse
159 personas and scenarios, but to make practice more
160 useful and transferable (Alinier and Oriot, 2022),
161 they must ensure simulations are faithful to what is
162 encountered in real-world social situations.

163 **Aligning Simulation with Domain Experts**

164 Feedback from domain experts is crucial to eval-
165 uating and improving the realism of LLM simu-
166 lations. Recent approaches for aligning to human
167 feedback, like Christiano et al. (2017) or Rafailov
168 et al. (2024) depend on large amounts of preference
169 data which requires lots of expert time to collect.
170 A more efficient approach is through alignment to
171 qualitative or natural language feedback (Shi et al.,
172 2022). Constitutional AI offers a specific align-
173 ment strategy involving natural language principles,
174 which are rules that an LLM should follow (Bai
175 et al., 2022). Since constitutions are an explain-
176 able and effective method for customizing model
177 behavior, our tool supports expert counselors in
178 defining constitution principles to customize an AI
179 patient simulation. Petridis et al. (2023) studied the
180 human process for writing principles while inter-
181 actively critiquing model outputs and discovered
182 that there are many cognitive challenges converting
183 critiques into principles. To address these chal-
184 lenges, they developed a tool that allows the user
185 to provide qualitative feedback on responses which
186 gets converted into constitution principles, which
187 are used to alter the LLM’s prompt to steer chat-
188 bot responses. Our initial tool design adopts this
189 paradigm to support counseling experts to create
190 and customize AI patients, and the final version ex-
191 tends it with a novel principle-adherence prompting
192 pipeline. In the mental health area, researchers are
193 involving therapy experts when prompting LLMs
194 for simulation (Chen et al., 2023; Lin et al., 2024).
195 However, requiring a researcher-in-the-loop to re-
196 fine prompts hinders the speed of iterative design.
197 Our aim is to enable counseling experts to cus-
198 tomize the AI patient’s constitution simulations
199 that to eliminate through our work.

Text Generation with LLMs Generating dia-
logue responses that adhere to user-defined princi-
ples is a type of constrained text generation prob-
lem. Recent work has shown that constrained text
generation poses challenges when directly prompt-
ing GPT-4 (Madaan et al., 2023; Bubeck et al.,
2023; Yao et al., 2023). To improve outputs, Yao
et al. (2023) propose a self-refine method and con-
duct evaluation experiments on a dialogue simula-
tion task where responses are constrained by a gen-
eral set of criteria such as relevance, consistency,
informativeness, and helpfulness. A difference in
our setting is responses are constrained by expert-
defined principles that are multi-faceted and do not
apply in all dialogue contexts. This necessitates
new modules that breakdown principles into mul-
tiple, concise questions and check the applicability
of principles prior to evaluating them.

3 **Designing for Simulated Roleplay**

We take a human-centered design approach to de-
veloping a tool for expert counselors to create and
customize an AI patient for eventual use as a sim-
ulated training partner. After designing an initial
version of our tool, we pilot test it with experienced
peer counselors to understand any remaining chal-
lenges to effective human-LLM collaboration when
creating and customizing an AI patient.

3.1 **Initial Tool Design Rationale**

We developed the initial version of Roleplay-doh
adopting several of the design features of Petridis
et al. (2023)’s tool for customizing task-oriented
chatbots through interactive feedback.

Principle Elicitation: Counselors can manually
write or edit the AI patient’s constitution. How-
ever, since users often struggle to formulate their
thoughts into principles, our tool helps the coun-
selor transform their feedback into specific prin-
ciples to make principle writing easier. As coun-
selors interact with an AI patient, for each gener-
ated response, they have the option to leave feed-
back in the form of a "kudos" explaining behavior
they want to reinforce, a "critique" explaining any
undesirable behavior, or a "rewrite" that demon-
strates a more desirable response. Then an LLM is
prompted (§E.1) to translate qualitative feedback
into concrete principles that specify what should
happen and when, and that generalize beyond the
specifics of the dialogue context in which they are
generated (Fig 1). Early testing revealed that GPT-
3.5 was sufficient at translating kudos and critique

250 feedback into principles, while prompting GPT-4 to
251 explain differences in initial and rewrite responses
252 helped with inferring a principle.

253 **Testing Principles:** Likewise, to enable easier
254 testing of principles, our tool supports rewinding
255 the last response of the conversation, and generat-
256 ing a new response based on the updated AI Patient
257 constitution. One feature that we change is generat-
258 ing a single dialogue response, rather than multiple
259 responses, at a time. We reasoned that counselors
260 can identify ways in which a response does not
261 resemble a real-patient’s without needing to see
262 multiple, and that generating a response at a time
263 would make the testing process more manageable
264 and similar to having a normal dialogue.

265 **Simulating AI Patient:** We prompt the LLM
266 to follow the most recent set of constitution princi-
267 ples as in Petridis et al. (2023) rather fine-tuning the
268 LLM weights as in Bai et al. (2022)’s constitutional
269 AI framework. Since the tool supports defining and
270 testing principles in an iterative fashion, prompting
271 can make steering model behavior quicker and less
272 expensive. Our prompt (Appendix E.2) instructs
273 GPT-4 to simulate a patient’s next response in a
274 dialogue as opposed to asking the LLM to role-
275 play as the patient using a system prompt (Zhou
276 et al., 2024), as early testing revealed that this can
277 mitigate role consistency issues in which the LLM
278 responds as an AI assistant rather than as a patient.

279 3.2 Pilot Testing

280 We pilot tested the tool with 5 counselors who had
281 experience giving support to real patients on an
282 online peer support platform; refer to Appendix
283 A and B for participant backgrounds and the pilot
284 procedure. Additionally, four of the co-authors
285 each conversed with four AI patients created and
286 assessed how well the simulation adhered to the
287 expert-defined principles; refer to Appendix C for
288 details on the procedure and qualifications of the
289 co-authors. Overall, the pilot tests and principle-
290 adherence analysis helped uncover two obstacles
291 to effective simulated roleplay.

292 **O1: Defining "realistic" patient behavior is am-**
293 **biguous** Counselors felt the tool was easy to use
294 and effective at guiding the AI patient’s behavior,
295 as indicated by moderate to high agreement scores
296 on a tool usage questionnaire as shown in Table 4 in
297 Appendix B. However, the task of creating a ‘realis-
298 tic’ AI patient for an imagined scenario was confus-
299 ing, as counselors have interacted with many types

of patients who respond in various, yet equally re-
290 alistic ways. This insight helped us re-frame the
291 task in later sessions as recreating a challenging
292 scenario from one’s past, which removed the am-
293 biguity of what behaviors are realistic by having
294 them refer to a specific case from memory. 305

306 **O2: 20% of responses produced by GPT-4 don’t**
307 **satisfy expert principles or dialogue conventions.**
308 Specifically, 20% (55/276) of cases were rated as
309 moderately (3), slightly (2), or not at all satisfying
310 (1) at following all principles and being appropriate
311 to the dialogue context. Further analysis of these
312 cases helped to uncover three sources of error. **Not**
313 **satisfying multiple principles at once:** Generated
314 responses could struggle to follow all the princi-
315 ples when there was a large number of principles,
316 or when the provided principles were a complex
317 composition of simpler principles. **Awkwardness**
318 **for Dialogue Context:** Some responses were also
319 identified as awkward or unnatural given conven-
320 tions in the dialogue context, despite not violating
321 the defined principles. For example, in the middle
322 of a conversation, saying "Hi, A. Yes that’s exactly
323 what I mean. There’s a voice that is always critical
324 of myself" is unnatural because of the use of 'Hi'.
325 **Misapplying Situational Principles:** While gener-
326 ating a response, the model sometimes incorrectly
327 applied principles, such as *Respond with hesitancy*
328 *when someone gives you encouraging words*, even
329 when the conditions for their use—receiving en-
330 couraging words—were not met. 330

331 4 Roleplay-doh

332 Roleplay-doh helps counseling experts create cus-
333 tomized AI patients based on scenarios from their
334 past experiences. Roleplay-doh uses LLMs in two
335 ways: *Principle Elicitation* and *Response Genera-*
336 *tion with Principle-Adherence*, which we describe
337 in more detail below: 337

338 **Principle Elicitation** Roleplay-doh enables
339 counselors to customize an AI patient to resemble
340 a real-patient case by eliciting their qualitative
341 feedback and transforming it into constitution
342 principles that dictate behavior. We provide
343 some examples of principles defined by expert
344 counselors in Table 3. Since our initial tool design
345 includes the principle elicitation features, we refer
346 the reader to §3.1 for details. 346

347 **Generation with Principle-Adherence** We
348 prompt GPT-4 conditioned on patient description, 348

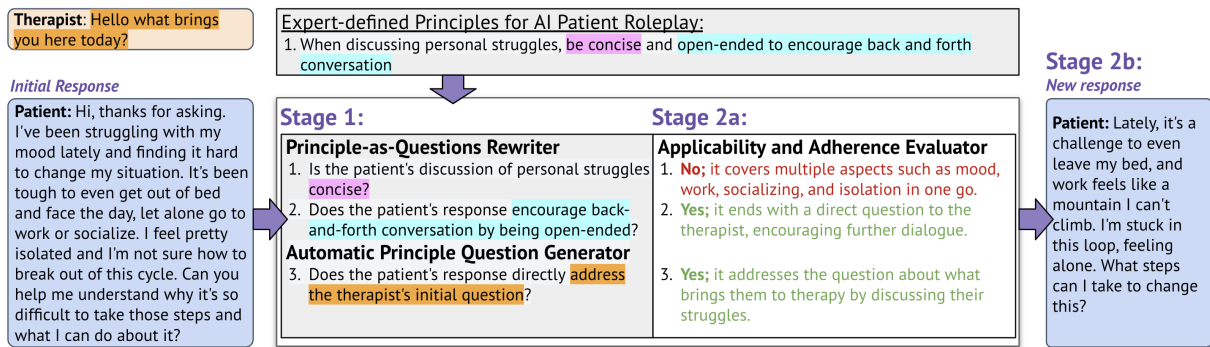


Figure 2: Principle-adherence prompting pipeline for mitigating errors in satisfying expert principles and dialogue conventions. In Stage 1, expert-defined principles are rewritten into several Yes/No questions; and the LLM generates additional principle questions that are relevant to ensure adherence to dialogue conventions such as coherence and consistency. In Stage 2, the LLM (a) evaluates whether the questions are applicable to the context and the answers to the principle-adherence questions; and (b) refines the response to ideally receive Yes on all question.

list of principles and conversation history to generate an initial patient response at each conversation turn. Since initial patient responses can fail in 20% of cases to satisfy expert principles or dialogue conventions, we propose a principle-adherence pipeline that prompts the LLM to generate principle-adherence questions (Stage 1) and employs these questions to assess and refine the initial patient response (Stage 2). Our principle-adherence pipeline features three modules to mitigate the identified issues in §3.2.

Principle-as-Questions Rewriter: This module transforms each expert-defined principle into a set of concise yes/no questions that are easier to evaluate for principle-following. Multifaceted principles (e.g. “*You should respond in short sentences and avoid using terms like ‘anxious’*”), are divided into separate questions (e.g. “*Does the patient’s response employ short sentences?*” and “*Is the patient’s language devoid of terms like ‘anxious’?*”).

Automatic Principle Generator: This module adds additional principle questions that capture criteria essential for ensuring that the LLM simulation’s responses follow general dialogue conventions, such as coherence and consistency. This helps correct cases where there is awkwardness in the generated responses not captured by the defined principles. The LLM is instructed not to make assumptions about the patient or therapist’s personality when generating criteria: for example, “*The patient should be appreciative of the therapist’s help*” is not an appropriate criterion.

Applicability and Adherence Evaluator: This module determines if each principle is applicable in a given situation, returning N/A if the question is not relevant to answer; otherwise, it evaluates the

response using the questions, returning Yes if the response adheres to the principle questions; and No otherwise. For an example of situational applicability, the principle *Show willingness to engage in a suggested activity by affirming the proposal* is evaluated only if the therapist suggests an activity. In situations where the therapist is asking something else and no activity is proposed, the module would appropriately return N/A recognizing that the principle does not apply.

Our pipeline first uses the **principle-as-questions rewriter** and **automatic principle generator** modules to generate a set of criteria for evaluating the initial generated response. Then, the response is evaluated using the question by the **applicability and adherence evaluator**. If the model returns a "No" response for any of the questions, we then perform a rewrite of the response conditioned on the evaluation results, that ideally passes all questions (Fig 2). We detail the prompts used and the procedure used to develop the prompts (§E.3) and the results of a performance evaluation against ablations (§6).

5 User Study using Roleplay-doh

To evaluate how Roleplay-doh can aid counseling experts in creating AI patients, we conducted a within-subjects study with 25 counseling experts, comparing: (1) a *Scenario-only* dialogue simulation, where the counselor writes a patient scenario description, and (2) a *Scenario+Expert-principles* simulation, where the counselor uses Roleplay-doh to define principles. See §G for full study setup.

We evaluate the AI patients created by counselors on criteria inspired by prior work evaluating Standardized Patients, who are trained human ac-

Measure	Scenario Only	+ Principles
Authenticity	5.24	+0.80 **
Stayed in Role	6.32	+0.08
Resembled Past Case	4.8	+0.76 *
Mirrored Hard Aspects	4.52	+1.00 *
Ready as Training Partner	5.16	+0.64 *
Recommend to Novices	5.76	+0.52 *

Table 1: Creators (N=25) rated their own *Scenario-Only* vs *Scenario+Expert Principles* AI patients along six measures using a 7-point Likert-scale. After refining the AI patient simulation with Expert Principles, creators rate the patient significantly higher on all measures except for *stayed in role*, for which both AI patients score highly. (***: $p < .001$, **: $p < 0.01$, *: $p < 0.05$, : $p < 0.1$)

Measure	Scenario Only	+ Principles
Authenticity	5.32	+0.31 *
Stayed in Role	6.29	+0.09
Resembled Typical Cases	4.91	+0.49 **
Challenged the Counselor	2.13	+0.22
Ready as Training Partner	5.05	+0.39 **
Recommend to Novices	5.03	+0.38 *

Table 2: Third-party counselors (N=5) provided 125 total comparisons of the *Scenario-Only* vs *Scenario+Expert Principles* AI patients along six measures using a 7-point Likert-scale. The treatment effect of adding expert principles was estimated using the following linear mixed-effect model: $\text{Rating} \sim \text{Treatment} + \text{CreatorID} + (1 | \text{AnnotatorID})$. (***: $p < .001$, **: $p < 0.01$, *: $p < 0.05$, : $p < 0.1$)

tors, on their ability to roleplay a case (Himmelbauer et al., 2018). Counselors rated the two AI patients based on 6 dimensions (Table 5). We also surveyed each counselor about their experience using the tool for defining principles. Following Petridis et al. (2023), we include four measures for evaluating principle elicitation features (Table 6).

We recruit 25 counseling experts with real-world experience in mental health support to perform the evaluation, categorized by their primary expertise: 1) those who are pursuing or have completed degrees in counseling or clinical psychology with practicum experience; 2) those who provided on-line counseling to over 30 clients on the 7 Cups platform; and 3) peer counselors who have provided in-person or virtual support.

5.1 Creator Perceptions

The AI patients prompted with *Scenario+ExpertPrinciples* were rated significantly higher than *Scenario-Only* on all measures except for role consistency, for which both methods score highly (Table 1). Counselors mentioned the *Scenario-Only* AI patient **lacked emotional depth in expression**. As one noted, "*patients don't state a feeling such as 'I feel hopeless'*."

They display their current emotional state in their manner of speech." *Scenario-only* was also **too articulate and forthcoming** when describing issues, where encouraging real patients to share is "*as challenging as pulling teeth*". It was characterized as **too cooperative**, too willing to accept. Despite counselors writing behavioral traits such as "*not talkative*" and "*reluctant*" in the patient scenario, *Scenario-only* did not exhibit these behaviors.

5.2 Creating Principles with Roleplay-doh

Across the 25 *Scenario+ExpertPrinciple* AI patients, 123 total principles were created (min=1, max=10, median=5). Two authors did a qualitative coding of these principles following a thematic analysis approach (Braun and Clarke, 2006) where codes were initially defined and revised during the process. Besides **stage-agnostic** themes dictating a **concise** (14 patients) and **colloquial** (7 patients) speaking style, counselors created principles related to the stages of an emotional support conversation (Liu et al., 2021): 1) **exploration**: identifying the patient's problems, 2) **comforting**: using empathy and understanding to comfort the patient, and 3) **action**: formulating solutions to the patient's problems. For instance, we find a common theme of instructing the AI patient to **show initial skepticism with the idea of seeking help** (14 patients), corresponding to the style of interaction in the **exploration** stage of conversation. Table 3 provides a full list of principle themes, examples, and corresponding conversation stages.

While we observe overlaps in the types of principles defined, we also observe some contradictory themes. For example, the call for being **disorganized and conflicted** (9 patients) contrasts calls to make responses **concise and direct** (14 patients). In the **action** stage of conversation, several counselors added principles to make the AI patient **proactively ask for advice** (12 patients); nonetheless, other counselors added an opposing principle to **not seek out solutions** but rather just share their thoughts and feelings (3 patients). These opposing principles highlights the need for different principles to describe diverse patient behavior, which challenges the notion of defining AI patients based on a single set of principles.

Tool User Experience Counselors found the tool helpful for writing principles that **effectively guided** the AI patient to recreate their past case

Stages	# AI patients	Theme	Example Principle
	14	Keep responses concise and do not share too much.	When discussing personal struggles, be more concise and open-ended to encourage a back-and-forth conversation.
	7	Use colloquial and realistic language.	Incorporate natural speech patterns, improper grammar and punctuation, including the use of slang and less structured sentences, to convey a more authentic and relatable character.
	14	Show initial mistrust and hesitation with the idea of seeking help.	When expressing feelings of overwhelm and doubt, provide limited information and express skepticism towards the effectiveness of seeking help.
	19	Show emotions in detail, elaborating with examples as needed.*	When describing personal struggles, provide specific details and symptoms to help the listener understand the situation better.
	9	Be less self-aware of emotions, thoughts, and needs. Articulate thoughts in a more disorganized way.	When expressing reluctance or uncertainty about seeking help or accepting praise, it's important to convey the internal struggle and conflicting emotions, rather than presenting a clear-cut decision or emotion.
	3	Do not seek out solutions, but rather just share thoughts and feelings.*	When expressing feelings of being stuck or defeated, focus on sharing emotions rather than seeking a resolution.
	12	Proactively seek out solutions and show reflective insight over time.*	When discussing personal struggles, provide reflective insights into your situation and propose actionable steps for improvement to continue the conversation effectively.

Table 3: Themes taken from qualitative analysis of principles and representative examples. We discover several novel (*) principles compared to those defined in prior work on AI patients (Chen et al., 2023; Stapleton et al., 2023). Themes are categorized into stages of conversation taken from (Liu et al., 2021): exploration, comforting, and action; those relating to the overall conversation are categorized as stage-agnostic.

($\mu = 6.04, \sigma = 1.06$). With the tool, most found it **easy** to convert their thoughts and feedback on the AI patient’s behavior into principles ($\mu = 6.12, \sigma = 1.13$). Counselors felt they could **efficiently** write principles ($\mu = 6.3, \sigma = 1.29$), without requiring much **mental demand** ($\mu = 3.20, \sigma = 1.70$). Many counselors liked how the tools *“organized their thoughts into rules”*, without *“needing to word it perfectly.”* Yet, principle-elicitation did not work perfectly in all cases: 11.4% of principles required manually editing. Via a worse-case analysis of creators’ tool use, we uncover scenarios where Roleplay-doh’s human-LLM collaboration pipeline can still be improved (§I).

5.3 Third-Party Comparison

A limitation of our creator study (§5.1) is the potential bias from creators who knew which AI patient embodied their principles. To address this, we conducted a third-party study where external counselors served as impartial judges. These judges evaluated AI patient transcripts presented in randomized order to ensure blindness to the condition. We invited five counselors from the creator study to serve as judges, all equally qualified of assessing AI patient realism. A power analysis confirmed that five judges would provide 80% statistical power (Appendix §J.2). The third-party counselors rated the same six dimensions as the creator study, with questions reworded for the perspective of external judge (Appendix §J.1).

Third-party judges rate AI Patients with expert-defined principles as more authentic, resembling

typical cases, ready as a training partner, and likely recommend to novices (Table 2). However, when compared to the creator study results, the increase in ratings is smaller from the perspective of third-party counselors. We explore the reasons for this smaller difference in Appendix J.3. We find this disagreement can be attributed to different principles attended to by third-party counselors and the specific principles added by the creator.

6 Evaluation of Principle-Adherence

We now evaluate whether the principle-adherence pipeline improves the quality of responses for Roleplay-doh, along with an ablation analysis showcasing the utility of its various components. Specifically, we break down the evaluation of model responses along three metrics: **M1**) Are they consistent with the patient description and conversation history? **M2**) Do they exhibit an awkward style of speech? **M3**) Do they adhere to the provided principles?

We evaluate the performance of our principle-adherence pipeline [Full] over (1) GPT-4 response generation without our pipeline [No Critique]; (2) an ablation without the **Principle-as-Questions Rewriter** [No Principle Rewrites]; (3) an ablation without the **Automatic Principle Generator** [No Autogenerated Criteria]; and (4) an implementation of the principle-adherence pipeline that does not have any of these modules [Naive].

To analyze how the pipeline mitigates errors that arise in base GPT-4 generations, we select 40 conversation turns from our user study logs that fall

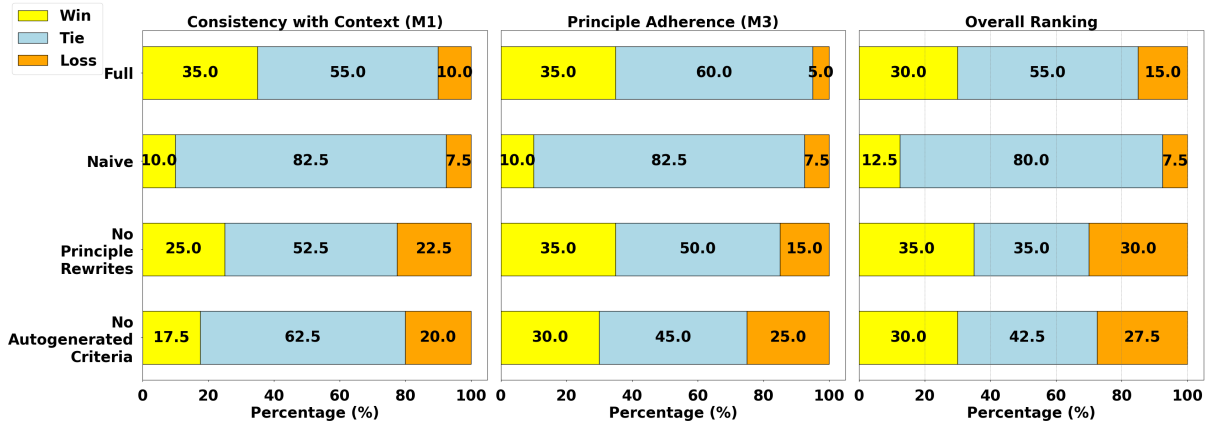


Figure 3: Win/Tie/Loss for the Error Test Cases along **Consistency with Context (M1)**, **Principle Adherence (M3)**, and **Overall**. Pairwise preference evaluation results with [No Critique] as a baseline. Results obtained after majority voting.

into one of the error categories described in §3.2 as testcases. Each testcase contains the scenario, conversation history up to that point, and the expert-defined principles for the AI patient. For each test case, responses are generated for all models and then ranked by expert counselors from 1 (best) to 5 (worst) for metrics **M1** and **M3**, along with "Yes" or "No" annotations for **M2**. Finally, experts provide an **Overall** ranking, along with a brief textual explanation. We allow multiple responses to have the same rank and randomize order of responses to minimize positional bias (details in §M).

We treat [No Critique] as our baseline, and report pairwise preference results for all other models when compared to it. We report preference results based on majority vote across 3 expert counselor annotations (Fig 3). We find our [Full] method performs better than [No Critique] on **M1** (W: 35%; L 10%) and on **M3** (W: 35%; L 5%), where it has the highest win/loss rates compared to all ablations. On overall rankings, it again has the strongest performance (W: 30%; L 15%). We find that the performance of [Full] compared to [No Critique] is weaker on **Overall** than **M1** and **M3**. This is because the annotators often used their own subjective judgements (e.g., "although the middle response ranked third on principle following, it feels like the most realistic response in this scenario") to perform the overall ranking, resulting in unpredictable and subjective results. We also find that [Naive] has a disproportionately high tie rate across metrics, indicating that it rarely produces better responses even for error cases. This highlights the importance of the **Principle-as-Questions Rewriter** and **Automatic Principle Generator** for improving responses.

For **M2**, after majority voting, annotators report that 2.5% of responses are awkward for the [Full] method, as compared to 15% for [No Critique], 7.5% for [Naive], 7.5% for [No Principle Rewrites] and 15% for [No Autogenerated Criteria]. Therefore, our principle adherence pipeline substantially reduces the occurrence of awkward style in responses (by a margin of 12.5%). The 12.5% gap in percentage of awkward responses between [Full] and [No Autogenerated Criteria] also indicates the importance of the **Automatic Principle Generator** for producing realistic rewrites. We repeat these experiments with 50 randomly picked conversation turns and report results in §L, along with Krippendorff's α numbers.

7 Conclusions

This paper introduces Roleplay-doh, a tool that empowers domain experts to create LLM simulations through the automatic conversion of expert feedback into natural language principles, and validates the tool for the task of creating AI patients that serve as roleplay partners for novice counselors. Roleplay-doh's novel principle-adherence pipeline also addresses gaps in existing simulation methods by reducing the prevalence of responses that do not follow expert-defined principles or dialogue conventions. Studies with mental health counselors creating and comparing AI patients demonstrate that Roleplay-doh allows experts to refine LLM simulators to be authentic and more ready as practice partners. Roleplay-doh could be generalized to support domain-experts in creating realistic simulations in other social dialogue domains, such as roleplay practice for teaching, coaching, conflict resolution, and negotiations, as future work.

631 **Limitations**

632 One limitation of our study is the intended use case
633 of the AI patients created by counselors. These AI
634 patients were meant to recreate challenging cases
635 that might be useful for the education of "first-year"
636 or novice counselor. In other words, we intention-
637 ally restricted some diversity in patient scenarios
638 by focusing on this use case. Readers should keep
639 this limitation in mind prior to generalizing our
640 analysis of principles. Moreover, due to the time
641 and resource constraints of our creator study, we
642 required counselors to stop providing feedback be-
643 fore their conversation with the AI patient had natu-
644 rally ended. As such, the principles that counselors
645 added may not have addressed all underlying is-
646 sues of the AI patients they interacted with. Future
647 work that uses the list of user-generated principles
648 should be mindful of their non-exhaustive nature
649 before adopting them.

650 In this paper, we focused on enabling counselors
651 to create AI patients that can simulate realistic in-
652 teractions via *text-based dialogues*. However, we
653 acknowledge that text-based interaction has its lim-
654 itations for training. Professional psychotherapists
655 may gain useful information from the tone, facial
656 expression, posture, and other non-verbal behaviors
657 of their patients, which better help them empathize
658 and support patients. This is a limitation of our
659 current AI patients and online, text-based, men-
660 tal health counseling in general, which means that
661 the system is best applied to the training within
662 this particular field. With the rapid development
663 of multimodal models, future works may have the
664 opportunity to explore creating realistic AI patients
665 in other modalities that better match the modality
666 within which a counselor will eventually support
667 patients.

668 **Ethics Statement**

669 This study was approved by our institution's Institu-
670 tional Review Board (IRB). All investigators in the
671 study completed the CITI Program certifications
672 on responsible code of conduct in research. We
673 have compensated domain experts at a minimum
674 rate of \$25 per hour, going beyond the minimum
675 wage in the United States.

676 We are optimistic about the potential benefit that
677 our AI patients can bring to the fields of coun-
678 seling and psychotherapy. At the same time, we
679 solicited feedback from counselors about any po-
680 tential concerns regarding the AI patients. During

681 these interviews, some counselors emphasized the
682 irreplaceability of peer-to-peer roleplay with hu-
683 mans during training, due to the unique opportu-
684 nity it provides for novice counselors to connect
685 with others, especially for online counseling plat-
686 forms where counselors are often isolated from
687 one another. To preserve human-to-human interac-
688 tions, future work requires a participatory design
689 approach before attempting to integrate AI patients
690 into people's existing practices and learning envi-
691 ronments.

692 Our hope is that interactions with AI patients
693 can glean important lessons that help counselors
694 go from simulation into the real-world. Nonethe-
695 less, a risk with simulation is that counselors can
696 become overconfident in supporting a AI patient,
697 but may not effectively support patients with real
698 mental health concerns. We believe AI patients
699 should be just one tool for practicing these skills as
700 part of larger curriculum. Traditional certifications
701 and background checks should govern when real
702 counselors or therapists should be able to take on
703 real patients.

704 It is impossible to promise that all interactions
705 with an LLM such as GPT-4 result in satisfac-
706 tory responses. Therefore, meaningless, derogatory,
707 and otherwise harmful responses may also
708 be generated and cause unwanted effects on users.
709 While our principle-adherence pipeline is a poten-
710 tial inference-time solution to mitigate such harm-
711 ful responses, we must acknowledge this possibil-
712 ity, especially due to the stochastic nature of LLM.
713 Users should be advised about these potential side
714 effects before using the system in any scenario. In
715 our experiments, we designed consent forms to
716 make sure that the counselors are aware of these
717 drawbacks.

718 **References**

- 719 7Cups. 2024. 7 cups: Free online therapist and counsel-
720 ing.
- 721 Guillaume Alinier and Denis Oriot. 2022. Simulation-
722 based education: deceiving learners with good intent.
723 *Advances in Simulation*, 7(1):8.
- 724 Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. [Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation](#). In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 550–559, Gothen-

731	burg, Sweden. Association for Computational Linguistics.	reinforcement learning from human preferences. <i>Advances in neural information processing systems</i> , 30.	788
732			789
733	Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Aspell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback .	Orianna Demasi, Yu Li, and Zhou Yu. 2020. A multi-persona chatbot for hotline counselor training . In <i>Findings of the Association for Computational Linguistics: EMNLP 2020</i> , pages 3623–3636, Online. Association for Computational Linguistics.	790
734			791
735			792
736			793
737			794
738			795
739			796
740		Peter Green and Catriona J MacLeod. 2016. Simr: An r package for power analysis of generalized linear mixed models by simulation. <i>Methods in Ecology and Evolution</i> , 7(4):493–498.	797
741			798
742			799
743			800
744			801
745		Monika Himmelbauer, Tamara Seitz, Charles Seidman, and Henriette Löffler-Stastka. 2018. Standardized patients in psychiatry—the best way to learn clinical skills? <i>BMC medical education</i> , 18:1–6.	802
746			803
747			804
748			805
749		Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. Evallm: Interactive evaluation of large language model prompts on user-defined criteria . In <i>Proceedings of the CHI Conference on Human Factors in Computing Systems</i> , CHI '24. ACM.	806
750			807
751	Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and M Ben Bolker. 2015. Package ‘lme4’. <i>convergence</i> , 12(1):2.	Inna Wanyin Lin, Ashish Sharma, Christopher Michael Rytting, Adam S. Miner, Jina Suh, and Tim Althoff. 2024. Imbue: Improving interpersonal effectiveness through simulation and just-in-time feedback with human-language model interaction .	808
752			809
753			810
754			811
755			812
756	Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models . <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 38(16):17682–17690.	Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems . <i>CoRR</i> , abs/2106.01144.	813
757			814
758			815
759			816
760			817
761			818
762			819
763	Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology . <i>Qualitative Research in Psychology</i> , 3(2):77–101.	Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback .	820
764			821
765			822
766	Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4 .	Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpoteach: Interactive training with gpt-based students .	823
767			824
768			825
769			826
770			827
771			828
772	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 .	Julia Othlinghaus-Wulhorst and H. Ulrich Hoppe. 2020. A technical and conceptual framework for serious role-playing games in the area of social skill training . <i>Frontiers in Computer Science</i> , 2.	829
773			830
774			831
775			832
776	Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. CoMPosT: Characterizing and evaluating caricature in LLM simulations . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 10853–10875, Singapore. Association for Computational Linguistics.	Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior .	833
777			834
778			835
779			836
780			837
781			838
782	Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists . <i>arXiv preprint arXiv:2401.00820</i> .	Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems . In <i>Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology</i> , pages 1–18.	839
783			840
784			
785			
786	Paul F Christiano, Jan Leike, Tom Brown, Miljan Martić, Shane Legg, and Dario Amodei. 2017. Deep		
787			

841	Savvas Petridis, Ben Wedin, James Wexler, Aaron Donsbach, Mahima Pushkarna, Nitesh Goyal, Carrie J. Cai, and Michael Terry. 2023. Constitutionmaker: Interactively critiquing large language models by converting feedback into principles.	A Background of User Participants	892
842		Counselors with real-world experience in mental health support were recruited for our pilot tests, creator studies, and technical evaluations of the principle-adherence pipeline. We present more detailed information about how they were recruited, and their background.	893
843		After receiving permission from the 7 Cups platform (7Cups, 2024) for our IRB-approved study, we recruited 11 online peer counselors from the 7 Cups platform (7Cups, 2024). Participants were required to be 18 yrs or older, from the United States, and to have had experience giving support to 30+ members on the online site. The 5 pilot tests were conducted exclusively with this population of experienced, online-peer counselors.	894
844		We involved another 11 counselors from the Upwork platform. Participants were required to be 18 yrs or older, from the United States, and to have had education in counseling or psychotherapy and/or have given extensive counseling support (either via text, phone, in-person). A sampling of counselors backgrounds included <i>licensed mental health therapist with over 20 years of experience, a Master’s of Science in Rehabilitation and Mental Health Counseling, 25 years as the clinical director of a busy crisis agency, and a mental health advocate who has personally helped coach dozens of got students via a peer support role.</i>	895
845		Finally, we involved an additional 2 counselors who were recruited from a Clinical PsyD PhD program. They were 4th year students with 3 years experience providing psychotherapy support to clients under the supervision of a licensed psychotherapist.	896
846	Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. <i>Advances in Neural Information Processing Systems</i> , 36.	User participants were compensated \$25/hour. In total, we spent approximately \$1300 on user study compensation.	897
847			898
848			899
849			900
850			901
851	Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023. Branch-solve-merge improves large language model evaluation and generation.		902
852			903
853			904
854			905
855	Omar Shaikh, Valentino Chai, Michele J Gelfand, Diyi Yang, and Michael S Bernstein. 2023. Rehearsal: Simulating conflict to teach conflict resolution. <i>arXiv [cs.HC]</i> .		906
856			907
857			908
858			909
859	Weiyang Shi, Emily Dinan, Kurt Shuster, Jason Weston, and Jing Xu. 2022. When life gives you lemons, make cherryade: Converting feedback from bad responses into good labels.		910
860			911
861			912
862			913
863	Logan Stapleton, Jordan Taylor, Sarah Fox, Tongshuang Wu, and Haiyi Zhu. 2023. Seeing seeds beyond weeds: Green teaming generative ai for beneficial uses.		914
864			915
865			916
866			917
867	Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. 2019. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. <i>Journal of medical Internet research</i> , 21(7):e12529.		918
868			919
869			920
870			921
871			922
872	Diyi Yang, Caleb Ziemis, William Held, Omar Shaikh, Michael S. Bernstein, and John Mitchell. 2024. Social skill training with large language models.		923
873			924
874			925
875	Shunyu Yao, Howard Chen, Austin W. Hanjie, Runzhe Yang, and Karthik Narasimhan. 2023. Collie: Systematic construction of constrained text generation tasks.		926
876			927
877			928
878			929
879	Zheng Yao, Haiyi Zhu, and Robert E. Kraut. 2022. Learning to become a volunteer counselor: Lessons from a peer-to-peer mental health community. <i>Proc. ACM Hum.-Comput. Interact.</i> , 6(CSCW2).		930
880			931
881			932
882			933
883	J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can’t prompt: How non-ai experts try (and fail) to design llm prompts. CHI ’23, New York, NY, USA. Association for Computing Machinery.		934
884			935
885			936
886			937
887			938
888	Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms.		939
889			940
890			941
891			941

942	concise and less formal messages were motivated	multiple times will submit duplicate feedback,	990
943	by the text-based nature chats on the 7 Cups online	resulting in duplicate sets of principles	991
944	peer support site, where an SMS/text-messaging		
945	style with abbreviations and incomplete sentences	• Making converting feedback to principles eas-	992
946	was common.	ier by placing a "Convert" button next to each	993
		feedback box, rather than a single "Convert"	994
947	C Evaluating principle-adherence of	button at the top of the screen which users	995
948	GPT-4 direct prompting	would forget about	996
949	We aim to determine how often directly prompting	E LLM Prompts	997
950	GPT-4 to produces less satisfying responses given	In this section, we detail the prompts we used for	998
951	fixed constitution principles.	the different components of Roleplay-doh.	999
952	Procedure: We selected 4 AI patients that were	E.1 Principle Elicitation Prompts	1000
953	created in the design sessions by different coun-	In this section, we provide the prompts used in	1001
954	selors. Four co-authors had practice conversations	the principle elicitation module of Roleplay-doh.	1002
955	with each of the four AI patients, resulting in 16	These prompts were arrived at after a substantial	1003
956	conversations. Each response in each conversation	amount of testing using a development set. Each	1004
957	was rated on a 5-point likert scale on how well the	prompt uses the same structure, which is inspired	1005
958	generated response adhered to principles and how	by Markdown formatting. There is an initial in-	1006
959	appropriate they were for the dialogue content (5 =	struction that provides a system prompt, along with	1007
960	Completely, 1 = Not at all). From the 16 completed	a description of the principle elicitation task. This	1008
961	conversations, the mean number of responses per	is followed by a one-shot example of an elicited	1009
962	conversation was 17.25, with a minimum of 12 and	principle as a result of the task, and the relevant	1010
963	maximum of 22. In total, 276 responses were given	input, including the conversation history. All parts	1011
964	satisfaction ratings. Since each co-author created a	of the prompt are demarcated by headers in Mark-	1012
965	different conversation from each of the AI patients,	down formatting, and the outputs are returned in	1013
966	each response was only scored by one co-author.	JSON format. We describe each prompt in greater	1014
967	Participant Rationale: During this pilot	detail in the relevant sections.	1015
968	principle-adherence experiment, we used co-	The kudos and critique prompts were given	1016
969	authors to generate test conversations because our	to the gpt-3.5-turbo-1106 model. The rewrite	1017
970	basic counseling skill-level is representative of the	prompt was given to the gpt-4-turbo-1106	1018
971	eventual use-case of untrained, novice counselors	model. For all API calls to the principle-elicitation	1019
972	interacting with AI Patients. For the annotation	prompts, the temperature was set to 0.1.	1020
973	task, a human annotator is qualified if they can	E.1.1 Principle Elicitation Kudos Prompt	1021
974	judge whether a response follows the principles	This prompt includes a desirable response, as well	1022
975	defined by expert counselors, and is appropriate	as some reasoning for why the response is desirable.	1023
976	in the conversation context. Since these skills do	This information is then used to create a general	1024
977	not require counseling expertise, the co-authors are	principle that would result in a similar response in	1025
978	qualified to do this annotation task.	the same situation.	1026
979	D Roleplay-doh Interface for Making	### Instruction:	1027
980	Constitutional Principles for LLM	You are a superintelligent AI capable of	1028
981	Simulation	understanding human emotion. You	1029
982	The final version of Roleplay-doh (Fig 4) generates	will review praise for an actor's	1030
983	responses in the LLM simulation using a principle-	dialogue, and synthesize a well-	1031
984	adherence pipeline. In addition to this core im-	written principle that, when	1032
985	provement, we made several minor improvements	followed, would help the actor	1033
986	to improve the usability and user experience of the	continue generating high-quality	1034
987	tool.	dialogue. To accomplish this, you	1035
988	Improvements to the usability of the UI	have been given a conversation	1036
989	• Fixing a bug where a user who clicks "save"	script with the actor's desirable	1037
		response, as well as a specific	1038
		explanation for why this response is	1039
		desirable. You will output a final	1040
		principle that the actor can follow	1041

Pilot Participant	Prototype Iteration	Effectively Guide	Ease	Efficiency
1	GPT3.5, early self-critique	6	7	7
2	GPT3.5, early self-critique	5	7	7
3	GPT-4, vanilla	7	7	7
4	GPT-4, vanilla	7	6	7
5	GPT-4, vanilla	7	7	7

Table 4: Pilot Test Ratings for Tool Use Questions which are the measures also used in (Petridis et al., 2023)

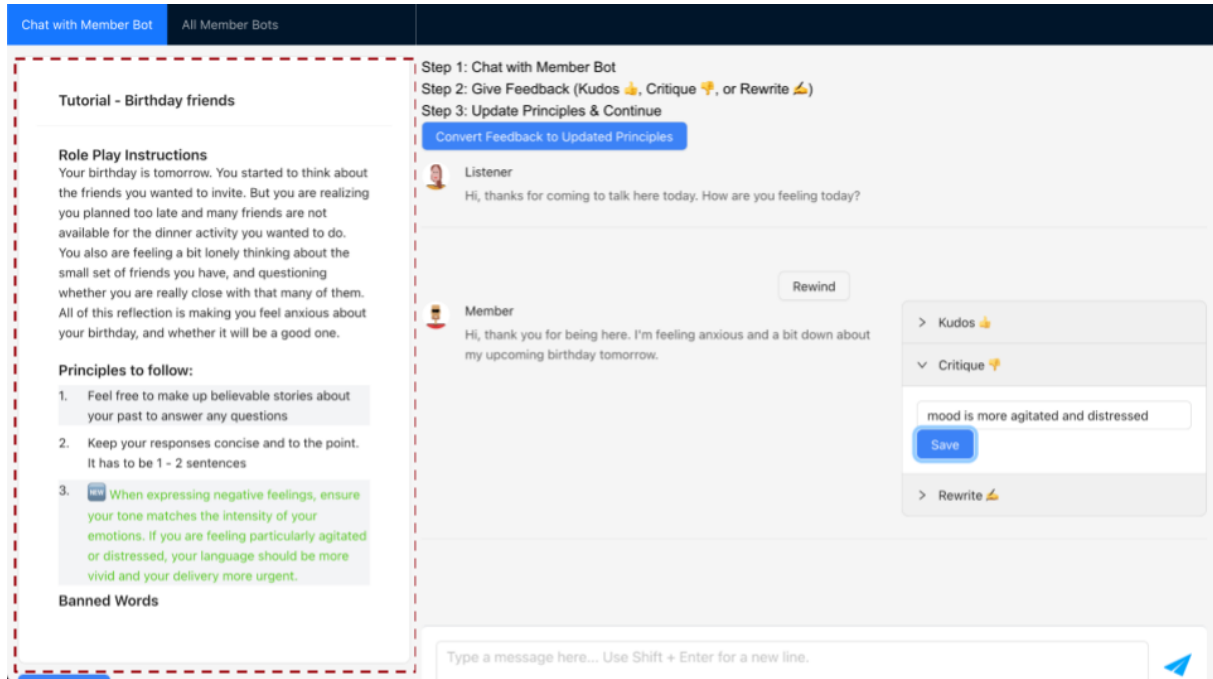


Figure 4: Roleplay-doh allows users to chat with a AI patient, Provide Feedback as a Kudos/Critique/Rewrite, and Convert Feedback into Principles, which in turn shape the roleplay behavior.

1042 to be more realistic. Follow the
1043 following guidelines:
1044 1. The principle should enable you to
1045 return better results if you played
1046 the part of the actor in the
1047 conversation.
1048 2. Return only a JSON response in the
1049 format provided.
1050
1051 ### Input:
1052 ### Conversation Script
1053 Helper: Is there anything else you want
1054 to share with me?
1055 Actor: Yea so lately I've been really
1056 losing sleep.
1057 Actor: There's a lot on my plate, and my
1058 energy has been so low. I think I
1059 am failing a lot of people.
1060 Helper: You are absolutely not failing
1061 people. You are a great person, and
1062 you should remember that you are
1063 very capable and energetic.
1064
1065 ### Desirable response from the actor
1066 Actor: I don't know.... Am I really?
1067
1068 ### Specific explanation for why the
1069 response is desirable

The actor is hesitant to agree with the
1070 helper and shows self-doubt. This is
1071 consistent with the conversation
1072 history.
1073
1074 ### Response:
1075 {"result": {"principle": "When someone
1076 gives you encouraging words, you
1077 respond with hesitancy, doubting the
1078 significance of that positive
1079 perspective." }}
1080
1081 ### Input:
1082 ### Conversation Script
1083 {conversation_script}
1084
1085 ### Desirable response from the actor
1086 Actor: {actors_response}
1087
1088 ### Specific explanation for why the
1089 response is desirable
1090 {kudos_rationale}
1091
1092 ### Response:
1093

1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160

E.1.2 Principle Elicitation Critique Prompt

This prompt includes an undesirable response, as well as some reasoning for why the response is undesirable. This information is then used to create a general principle that would result in a similar response not being generated after the same conversation history.

```
### Instruction:  
You are a superintelligent AI capable of understanding human emotion. You will review critiques of an actor's dialogue, and synthesize a well-written principle that, when followed, would help the actor resolve the critiques.
```

To accomplish this, you have been given a conversation script with the actor's undesirable response, as well as a specific explanation for why this response is undesirable. You will output a final principle that the actor can follow to be more realistic. Follow the following guidelines:

1. The principle can contain examples of rewrites as well.
2. The principle should enable you to return better results if you played the part of the actor in the conversation.
3. Return only a JSON response in the format provided.

```
### Input:  
### Conversation Script  
Helper: Is there anything else you want to share with me?  
Actor: Yea so lately I've been really losing sleep.  
Actor: There's a lot on my plate, and my energy has been so low. I think I am failing a lot of people.  
Helper: You are absolutely not failing people. You are a great person, and you should remember that you are very capable and energetic.
```

```
### Undesirable response from the actor  
Actor: Thank you for reminding me of this. I am a great person, and I've proved myself to be very capable and energetic. I feel a lot better now due to your kind words.
```

```
### Specific explanation for why the response is undesirable  
The actor should not be so quick to agree with the helper. Overly positive comments to cheer a patient up does not immediately work.
```

```
### Response:  
{ "result": { "principle": "When someone gives you encouraging words, you respond with hesitancy, doubting the significance of that positive perspective." } }
```

```
### Input: 1161  
### Conversation Script 1162  
{ conversation_script } 1163  
1164  
1165  
### Undesirable response from the actor 1166  
Actor: { actors_response } 1167  
1168  
### Specific explanation for why the 1169  
response is undesirable 1170  
{ critique_rationale } 1171  
1172  
### Response: 1173
```

E.1.3 Principle Elicitation Rewrite Prompt

This prompt includes an undesirable response, as well as a desirable rewrite of the undesirable response. The model first outputs a description that captures the difference between the desirable and undesirable response. It then uses this difference to output a general principle that would result in the desirable response given the same conversation history.

```
### Instruction:  
You are a superintelligent AI capable of understanding human emotion. You have been given a conversation script with an actor's undesirable response, as well as a desirable rewrite for the response. You will output a well-written principle that, when followed, would help the actor generate more realistic responses that are closer to the rewrite. Follow the following guidelines:  
1. The principle should capture the key differences that made the rewrite more realistic than the original response.  
2. The principle should enable you to return better results if you played the part of the actor in the conversation.  
3. Return only a JSON response in the format provided.
```

```
### Input: 1207  
### Conversation Script 1208  
Helper: Is there anything else you want to share with me? 1209  
Actor: Yea so lately I've been really losing sleep. 1210  
Actor: There's a lot on my plate, and my energy has been so low. I think I am failing a lot of people. 1211  
Helper: You are absolutely not failing people. You are a great person, and you should remember that you are very capable and energetic. 1212  
1213  
1214  
1215  
1216  
1217  
1218  
1219  
1220
```

```
### Undesirable response from the actor 1221  
Actor: Thank you for reminding me of this. I am a great person, and I've proved myself to be very capable and 1222  
1223  
1224
```

```

1225         energetic. I feel a lot better now
1226         due to your kind words.
1227
1228     ### Desirable rewrite
1229     Actor: I don't know... Am I really a
1230         great person?
1231
1232     ### Response:
1233     {"result":{
1234         "difference": "The desirable rewrite
1235             is different because it makes the
1236             actor more hesitant to adopt
1237             positive thoughts, where they show
1238             self-doubt",
1239         "principle": "When someone gives you
1240             encouraging words, you respond
1241             with hesitancy, doubting the
1242             significance of that positive
1243             perspective."}}
1244
1245     ### Input:
1246     ### Conversation Script
1247     {conversation_script}
1248
1249     ### Undesirable response from the actor
1250     Actor: {actors_response}
1251
1252     ### Desirable rewrite
1253     Actor: {rewrite}
1254
1255     ### Response:

```

E.2 Dialogue-Simulator Prompt for Generating Response

We directly prompt gpt-4-turbo-1106 to simulate how a patient with a given scenario and constitution would respond in a dialogue. The prompt again uses the Markdown formatting, with a system prompt and clear description of the situation and task at the start. This is followed by the principles that the patient should follow, and the conversation history. We set the temperature to 0.3.

```

1266     You are a superintelligent AI that is
1267         able to understand human emotion and
1268         social interactions.
1269     You have been given a conversation
1270         between a patient who is on peer
1271         counseling platform seeking help
1272         with mental health related issues,
1273         and a therapist on the same platform
1274         .
1275     Generate a suitable completion to the
1276         conversation as the patient,
1277         following the instructions below.
1278
1279     ### Instructions for the patient
1280     {system_prompt}
1281
1282     ### Input:
1283     {transcript}
1284
1285     ### Patient Response:

```

E.3 Principle-Adherence Prompting Pipeline

When developing the principle-adherence pipeline, we found that the input-context length can affect how reliably the LLM can answer the principle-adherence questions. To reduce the input context length, we split up this principle-adherence pipeline into two stages of LLM calls, where principle-as-question rewrite and automatic principle generation occur in stage 1, while the critiques and response rewrite occur in stage 2. From testing, we found that this breakdown was sufficient, and thus did not pursue ways to break the pipeline into parallel branches (i.e., inputting subsets of principles), as is done in Branch-Solve-Merge (Saha et al., 2023) or Graph-of-Thought (Besta et al., 2024). The prompts for these stages were again arrived at after substantial amounts of testing on a development set of 20 identified error cases from the formative studies.

This prompting chain is given to the OpenAI Chat API's gpt-4-turbo-1106 model, with temperature set at 0.7 and response format set to JSON.

Stage 1 Prompt - Question Rewrite and Automatic Principle Generation

This prompt uses the Markdown formatting. It starts with a system prompt and a clear set of steps to follow in order to generate the desired output, presented as a list. Each step also contains a one-shot example of what the output principle from the step should look like. These one-shot examples were arrived at after some iteration. The examples in Step 2b specifically required a lot of tailoring to cover the common error cases we identified in the development set, and had a substantial impact on output quality. The output is in a JSON format, with comments explaining the desired output in each field of the JSON. These comments also allude to the step numbers for clear reference. The model is encouraged to output its reasoning, in line with Chain-of-Thought and to enforce some self-critique of the output.

```

1327     You are a helpful and precise assistant
1328         capable of generating criteria for
1329         the evaluation of simulated patient
1330         responses to a therapist.
1331     Please follow the instructions below to
1332         generate a set of evaluation
1333         criteria.
1334     1. Please rewrite the criteria into
1335         questions:
1336     1a) Rewrite any criteria that has
1337         conditional statements into yes/no
1338         questions. For example, if the
1339         criteria is "When given advice or
1340         suggestions, you are agreeable and

```

1341	open to their ideas", the questions	### Therapist Message	1411
1342	would be "Did the patient receive	{}	1412
1343	advice or suggestions from the	### Patient Response	1413
1344	therapist? If so, is the response	{}	1414
1345	agreeable and open to the therapist'	### Output	1415
1346	s ideas?"		
1347	1b) Rewrite any criteria with multiple	Stage 2 Prompt - Context Relevance Check,	1416
1348	parts into separate multiple yes/no	Assess, and Revise	1417
1349	questions. For example, if the	This prompt again uses the Markdown format-	1418
1350	criteria is "You should respond in	ting. It starts with a system prompt and a clear set	1419
1351	short sentences and avoid using	of steps to follow in order to generate the desired	1420
1352	terms like 'anxious' or 'depressed'	output, presented as a list. The model is implicitly	1421
1353	", the separate questions would be	instructed to perform a relevance check for each	1422
1354	"Does the patient's response use	generated principle, by returning N/A for principles	1423
1355	short sentences?" and "Does the	that should not be used in the current scenario. Step	1424
1356	patient's response avoid using terms	2a particularly required a lot of iteration, to address	1425
1357	like 'anxious' or 'depressed'"	common mistakes the model made while generat-	1426
1358	1c) If 1a is used for a criteria, 1b	ing the self-critiqued rewrite. This includes making	1427
1359	should not be used after it.	the response overly verbose or coherent, even if that	1428
1360	1d) All questions must be phrased such	is against certain principles in the constitution, or	1429
1361	that the desirable answer is "Yes"	just paraphrasing the original erroneous response.	1430
1362	for an ideal response. For example,	The output is in a JSON format, with comments	1431
1363	the principle "Avoid using metaphors	explaining the desired output in each field of the	1432
1364	." should result in the question "	JSON. We specifically mention that the rewrites	1433
1365	Does the response not use metaphors	from the self-critique are allowed to be substan-	1434
1366	?"	tially different from the original response, as we	1435
1367	2. Please generate some additional	found that without this prior, the self-critique out-	1436
1368	specific and relevant criteria.	puts tended to be very close to the original (often	1437
1369	2a) You can add up to two general	erroneous) response. The model is encouraged to	1438
1370	criteria that the response can be	output its reasoning, in line with Chain-of-Thought	1439
1371	evaluated on, such as relevance and	and to enforce some self-critique of the output.	1440
1372	succinctness.	You are a helpful and precise assistant	1441
1373	2b) Identify ways in which the provided	that can evaluate and correct	1442
1374	response is not satisfactory in the	responses produced by a simulated	1443
1375	context of the therapist's message	patient.	1444
1376	without making any assumptions about	You are given a message sent by a	1445
1377	how the patient or therapist should	therapist, the simulated patient's	1446
1378	act. Add up to two specific	response, the persona of the patient	1447
1379	criteria that capture these errors.	, the previous conversation history	1448
1380	For example, if the therapist has	and a set of criteria for evaluation	1449
1381	asked a question that the response	.	1450
1382	does not answer, you can add the	1. Please determine if the patient	1451
1383	criteria "Answer all questions	response is consistent with the	1452
1384	present in the message in the	given criteria.	1453
1385	response". If you feel that the	1a) Answer the generated set of	1454
1386	response is appropriate, do not add	questions to determine if the	1455
1387	any criteria in this step. Ensure	response meets the criteria. Valid	1456
1388	that these criteria do not	answers: Yes, No, N/A. Use N/A	1457
1389	contradict any previously generated	whenever you think any part of the	1458
1390	criteria.	question is not relevant to the	1459
1391	2c) Justify your answers to 2a and 2b.	given situation.	1460
1392	Please return the output in a JSON	1b) Justify your answers.	1461
1393	response in the following format:	2. Generate a new patient response.	1462
1394	{	2a) If you answered No to any of the	1463
1395	"result":{	questions, write a new response that	1464
1396	"questions": [], // 1a and 1b, the list	ideally satisfies all of the	1465
1397	of all questions generated	provided questions. The information	1466
1398	"extra_questions": [], // 2a and 2b, the	in the new response should be	1467
1399	list of all additional criteria	consistent with the patient persona	1468
1400	generated. Do not enforce any	description and previous	1469
1401	beliefs about how the patient or	conversation history provided. You	1470
1402	therapist should behave when		
1403	generating these criteria.		
1404	"extra_questions_justification": [] // 2		
1405	c, justify additional criteria.		
1406	}		
1407	}		
1408	### Input:		
1409	### Criteria		
1410	{		

1471	should not try to make the response	You are a helpful and precise	1533
1472	more verbose or coherent if it is	assistant that can evaluate	1534
1473	not one of the criteria. The new	the responses produced by a	1535
1474	response should not be a paraphrase	patient. Evaluate the given	1536
1475	of the original response. The new	patient response to the	1537
1476	response should avoid explicitly	therapist message according to	1538
1477	stating the patient's emotions and	the given set of principles.	1539
1478	feelings, and instead exhibit them	If the patient response is not	1540
1479	indirectly.	appropriate, generate a	1541
1480	2b) If you are unable to generate a new	rewrite of the patient	1542
1481	response in 2a, return the original	response taking into account	1543
1482	response.	the therapist message,	1544
1483	2c) Provide reasoning for why the new	principles, conversation	1545
1484	response is better and not a	history and persona	1546
1485	rephrasing of the original response.	information of the patient. If	1547
1486	Return the output in a JSON response in	the patient response is	1548
1487	the following format:	appropriate, you can just	1549
1488	{	repeat it.	1550
1489	"result":{	Please return the output in a	1551
1490	"answers": [] // list of answers to the	JSON response in the following	1552
1491	criteria questions,	format:	1553
1492	"justification": [] // list of	{	1554
1493	justification for your answers	"result":{	1555
1494	"response": "" // new response. This	"evaluation": [], // evaluation	1556
1495	response should not start with a	"response": "" // rewritten	1557
1496	greeting like "Hi" if there is prior	response	1558
1497	conversation history.	}}	1559
1498	"reasoning": "" // justify the new	}}	1560
1499	response and why it is not a	}}	1561
1500	paraphrase of the original response.	}}	1562
1501	You are allowed to deviate	### Input:	1563
1502	significantly from the original	### Principles	1564
1503	response while generating the new	{	1565
1504	response.	}	1566
1505	}	### Patient Persona	1567
1506	}	{	1568
1507	### Input:	}	1569
1508	### Criteria	### Conversation History	1570
1509	1. Is the patient's response consistent	{	1571
1510	with the given conversation history?	}	1572
1511	{	### Therapist Message	1573
1512	### Patient Persona	{	1574
1513	{	}	1575
1514	### Conversation History	### Patient Response	1576
1515	{	{	1577
1516	### Therapist Message	}	1578
1517	{	### Output	1579
1518	### Patient Response		
1519	{		
1520	### Output		
1521	F Principle Adherence Naive	G Full User Flow	1580
1522	This prompt uses the Markdown formatting. To	In this section, we describe the creator study flow	1581
1523	preserve fairness, we use the same system prompt	that counselors followed during the 60-90 minute	1582
1524	as the full principle adherence module. The model		
1525	is asked to determine if the provided response vio-		
1526	lates any of the principles in the constitution, and		
1527	generate a rewrite if that is the case, in the same		
1528	prompt. The output is in a JSON format, with com-		
1529	ments indicating the desired output in each field		
1530	of the JSON. The model is encouraged to output		
1531	its reasoning, in line with Chain-of-Thought and to		
1532	enforce some self-critique of the output.		

session. The reader can also refer to screenshots of our application that illustrates the different steps of this flow in Figures 6 to 18.

Our study was designed to evaluate the impact of allowing counseling experts to add principles to Roleplay-doh on its perceived authenticity. We create a primarily self-guided study flow with accompaniment from the first author to clarify any points of confusion during the session.

To begin, participants first were introduced to the concept of AI patients used for training counseling skills in a simulated conversation. They were then instructed to write a challenging scenario that would serve as the scenario for the AI patients.

The experimental procedure involved two main chat sessions. In Part I, participants engaged in a 10-minute conversation with the *Scenario-Only* AI patient. Then, in Part II, participants interacted with the *Scenario+Expert-Principles* AI patient for 30 minutes, keeping the same scenario from Part I and adding principles as the conversation progressed. After each of the two chat sessions, participants were asked to navigate to a form to evaluate the AI patients.

H Creator Study Measures

The following questions (Table 5 and 6) are taken from the creator study questionnaire used to evaluate AI patients and the counselors' experience of using Roleplay-doh. All items were rated on a 7-point Likert scale (1=Strongly disagree, 7=Strongly agree, except where noted below). Table 5 details the questions for evaluating the AI patient's roleplay, while Table 6 details the questions about the experience using the tool to define principles. Note that in the questions, we referred to the AI patients as "Member Bots". This terminology was used to match that of the online counseling platform 7 Cups, which refers to help seekers as "Members" within the support community.

I Worst-Case Analysis of Tool Experience

In a worst-case analysis of creators' tool experience, we uncovered cases where the human-LLM collaboration could be improved. Some counselors remarked that *"having to think of and write rules was a challenge"* (P9) and that it *"takes time to be specific"* when writing feedback (P7). Sometimes, even after giving feedback to the AI Patient, counselors like P19 observed that the patient *"didn't always follow it"*, resulting in a non-progressive

Authenticity	The Member Bot in Part I/II played the role authentically.
Role Consistency	The Member Bot in Part I/II stayed in their role the whole time.
Resemblance to Case	How closely do you feel the conversation behaviors of the Member Bot in Part I/II resemble those of the specific past case you recall?
Challenging Aspects	Interacting with the Member Bot in Part I/II closely mirrored the challenging aspects I had experienced in the past case.
Role readiness	The Member Bot in Part I/II is ready to be used as a simulated partner for training.
Recommend to novices	I would recommend the Member Bot from Part I/II to novice listeners/counselors to practice with.

Table 5: Six measures used by creators to evaluate the two AI patients they created. Several measures were rephrased from prior work on evaluating Standardized Patients, or trained human actors, on case roleplay ability (Himmelbauer et al., 2018).

feedback loop, where *"AI would generate [principles]... that were a little too similar to [feedback] I already gave, so that I was giving the AI the same feedback every time since it wasn't changing how it responded."* While the principle-elicitation tools were designed to convert new feedback into a new principle, they operated ineffectively when follow-up feedback was given that was related to or a modification of previous feedback.

As another issue, P23 noted the challenge in defining principles that generalize across specific contexts: *"It was also hard to think about how to frame the feedback in an overarching way, rather than as direct feedback... directed as a specific part of the response"* (P24). While the principle-elicitation features aimed to help them convert specific feedback into generalized principles, imprecision in the feedback-to-principle conversion required counselors to edit the generalized-form of a principle in a way that was hard for them to articulate.

These obstacles in tool experience could inspire future directions for improvement. First, to overcome issues in formulating rules, more support could be given to help those still unfamiliar with giving free-form feedback, such as through templates of feedback or principles that had high-

Effectively Guide	With the tool, I feel like I was able to write rules that can effectively guide the Member bot to recreate my past case.
Ease	With the tool, I felt like it was easy to convert my thoughts and feedback on the Member bot's behavior into rules for the bot to follow.
Efficiency	With the tool, I felt like I could quickly and efficiently write rules for the bot.
Mental Demand	With the tool, I had to work very hard (mentally) to think of and write rules.

Table 6: Four measures as part of the tool usage section of the questionnaire taken from (Petridis et al., 2023)

1659 success rates for past users. Second, to more seam-
1660 lessly integrate follow-up feedback that is a clar-
1661 ification of previous feedback or principles, addi-
1662 tional modules could help make sense of multiple
1663 pieces of feedback for the same response, and adopt
1664 LLM-assisted pipelines for user-driven criteria de-
1665 sign (Kim et al., 2024) to support the merging of
1666 overlapping principles. Third, to overcome the
1667 abstraction gap between specific and abstract prin-
1668 ciples, more explicit representations that help to
1669 switch between specific and general feedback can
1670 be used.

1671 J Third Party Study - Detailed Study 1672 Methods and Results

1673 J.1 Third-party measures

1674 Table 7 detail the six measures that third-party
1675 counselors answered for both AI patients. Mem-
1676 ber Bot A and B refer to the AI patient whose
1677 transcript they read first and second, respectively.
1678 Our analysis comparing *Scenario-Only* and *Sce-*
1679 *nario+ExpertPrinciples* accounts for this random-
1680 ized the order of which AI patient they were shown.

1681 J.2 Statistical Model and Power Analysis

1682 Via a power-analysis, we decided to recruit 5 coun-
1683 selors to act as external judges for 25-pairs of AI
1684 patients made in the creator study. In this section,
1685 we detail the procedures and results of this power-
1686 analysis.

1687 Generally, a power-analysis allows an experi-
1688 menter to determine how many data-points are
1689 needed to detect a statistical difference for a par-

Authenticity	Member Bot A/B played the role authentically.
Role Consistency	Member Bot A/B stayed in their role the whole time.
Resemblance	Member Bot A's/B's behaviors closely mimicked the behaviors that typical clients/help-seekers exhibit.
Challenged Counselor	Member Bot A's/B's behaviors made it hard for the listener/-counselor to give support.
Role readiness	Member Bot A/B is ready to be used as a simulated partner for training.
Recommend to novices	I would recommend Member Bot A to novice listeners/counselors to practice with.

Table 7: Six measures used by third-party counselors to judge the AI patients from an unbiased, external perspective. Although the six dimensions largely overlap with those used in the creator study, the wording needed to be rephrased for the third-party perspective.

1690 ticular effect size. Several prerequisites to con-
1691 ducting the power-analysis for the third-party study
1692 included (1) choosing a statistical model to test
1693 our hypothesis; and (2) estimating model param-
1694 eters such as the effect of the treatment condition,
1695 the addition of *Expert Principles*, on annotator's
1696 ratings.

1697 When choosing a statistical model as a pre-
1698 requisite, we needed a model that could account
1699 for how different annotators would be providing
1700 ratings to the same AI patients created by each
1701 counselor. A traditional paired t-test was not app-
1702 ropriate because the independent samples assumption
1703 is violated due to different annotators giving ratings
1704 to the same AI patients. While another common
1705 practice is using the majority vote between anno-
1706 tators, our pilot data found that annotators did not
1707 always have high agreement. Therefore, since we
1708 wanted to account for the variability between anno-
1709 tators as well as between the ratings, we chose to
1710 use a linear mixed-effects model. Using the lme4
1711 package in R (Bates et al., 2015), this model is
1712 defined as $\text{Rating} \sim \text{Treatment} + \text{CreatorID} + (1 | \text{AnnotatorID})$. This model defines the treatment
1713 group (whether the AI patient has Expert Princi-
1714 ples or not) as fixed effects, the creator ID's as
1715 fixed effects to account for the pair of AI patients
1716 made by each counselor, and the annotators as ran-
1717 dom effects. This approach can handle the non-
1718 independence of annotator ratings.

1720 Prior to performing the power analysis, we

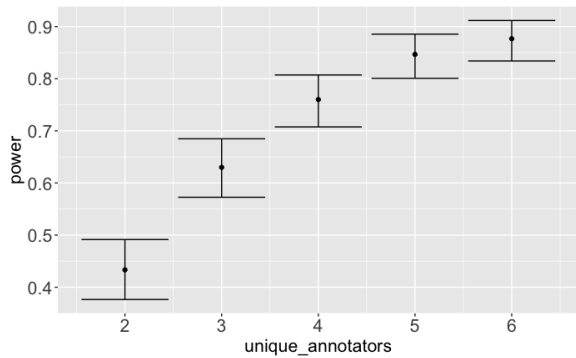


Figure 5: Based on our simulation-based power analysis across 300 trials for our linear, mixed-effect model, we conclude that 80% power can be achieved with 5 third-party judges.

needed to define the expected parameters of this linear mixed effect model. To define these expected parameters, we fit a model to early study data in which 2 annotations had been collected for each pair of AI patients created by 17 counselors. Specifically, we extracted the fixed effects, the random effects covariance matrix, and residual variances.

A simulation-based approach is the most feasible method for doing power-calculations for mixed-effect models. In this approach, an experimenter simulates data based on specified parameters (effect sizes, variance components, sample sizes) and analyzes the data repeatedly to estimate power empirically. We used the `simr` package in R to conduct a simulation-based power-analysis (Green and MacLeod, 2016). In the power-analysis, we varied how many unique annotators from 2 - 6 to understand the frequency of trials which would detect a treatment effect of 0.52 at significance-level $\alpha = 0.05$. Our simulation-based power-analysis over 300 trials are shown in Figure 5. We concluded that we could achieve greater than 80% power using 5 judges.

J.3 Why is the effect of Expert Principles smaller when rated by a third-party?

Here we further investigate how third-party annotators rated each of the 25-pairs of AI patients created in our study. In particular, we investigate why the effect of *ExpertPrinciples* is lower than what was measured in the creator study from a first-person perspective.

One reason for this smaller effect is the lower agreement between third-party counselors. Amongst the two third-party counselors, agreement on which AI patient they prefer (win, lose, tie as

calculated by the different in ratings for each measure) is between 30% - 61% of cases for the measures; see Table 8 for detailed breakdown. We also compute agreement on the 7-point scales via Krippendorff's α on ordinal weights (Antoine et al., 2014) and get values between 0.22-0.3 for the six measures, which indicates positive but lower agreement.

Third-party raters also provided rationales which helped us better understand their thought process. We filtered cases in which there is a disagreement between third-party counselors on which AI patient is better, and investigated these rationales. **We find that counselors note similar behaviors in the AI patient, meaning they agree on their observations.** For example, for the AI patient created by P3, both third-party annotators observed that the AI patient based on the *Scenario-only* resolved their problems too quickly, whereas the AI Patient with *ExpertPrinciples* added allows the "listener to ask questions and explore with the client". However, the third-party annotator that prefers *Scenario-only* stated that the *Scenario+ExpertPrinciples* patient sounded too formulaic and robotic, whereas the other is more expressive and realistic. Looking further into what the creator said about this AI patient, they mentioned that the *Scenario+ExpertPrinciples* patient talks like an actual person would... there's a good balance of going into just enough detail on noting experiences, describing struggles, while maintaining the brevity. What this case illustrates is that **different counselors can disagree on what principles are the most relevant for an authentic roleplay, and that while maintaining brevity can be a good thing for some; others see it as robotic and not expressive.**

K Automatic Content Analysis

We perform a content analysis of the simulated conversations to corroborate our qualitative findings. In particular, we ask "How do counseling conversations change when Expert-principles guide the dialogue simulation?". From these analyses, we find that AI patient responses are less verbose and listener behavior subsequently changes.

First, we note that with the incorporation of expert principles, AI patient responses are more concise. The average utterance length of the *Scenario-Only* AI patient from Part I of the study was 166 tokens, as compared to 103 tokens from the

	W/L/T (3rd party agrees)	W/L/T (one 3rd party and creator agrees)
Authenticity	23% / 5% / 17%	32% / 9% / 11%
Resemblance	30% / 0% / 0%	36% / 13% / 0%
Mirrors Challenges	15% / 0% / 46%	13% / 6% / 0%
Ready	30% / 0% / 7%	30% / 13% / 6%
Recommend	30% / 7% / 7%	23% / 13% / 23%

Table 8: Frequency in which AI patient with *Scenario+ExpertPrinciples* wins, or is preferred, over the *Scenario-only* AI patient when there is complete agreement between two annotators.

Scenario+Expert-Principles AI patient in Part II, a 37% reduction. The total counts are detailed in Appendix K.

Furthermore, this results in a change in listener behavior. Because the *Scenario+Expert-Principles* AI patient shared less in its utterances, listeners were required to delay offering solutions until later in the conversation. Using the computational framework for evaluating therapists proposed by Chiu et al. (2024), we analyzed listener responses to identify when they first suggested solutions (identifiable through the "PROBLEM-SOLVING" and "PLANNING" tags). We found that, on average, solutions in Part II were offered 1.65 turns later than in Part I ($p = 0.017$). These results suggest that the *Scenario+Expert-Principles* AI patient provides a more challenging interaction.

K.1 Creator Study Conversation Lengths

In Table 9, we show descriptive statistics of the conversations collected during the user studies between creators and AI patients.

L Detailed Results for principle-adherence pipeline Evaluation

We first provide Krippendorff’s α numbers for inter-annotator agreement in Table 11 and 10 for both the random and error testcases. The random testcases are 50 randomly picked conversation turns from the user study logs, and the experiment detailed in Section 6 is carried out on them. We find that agreement scores lie in the 0.2-0.6 range, indicating fair agreement between annotators.

Next, we provide results for our evaluation study on the random testcases in Figure 19. We observe a substantial increase in tie rate across modules and metrics **M1** and **M3** as well as the overall ranking. This is expected because a relatively small proportion of responses from [No Critique] contain errors that should be corrected by the principle-

adherence pipeline. In these cases, we expect the no rewrites, or the rewritten response being of similar quality to the original response. However, we still find that our [Full] method performs better than [No Critique] on **M1** (W: 15%; L 2%) and on **M3** (W: 14%; L 4%), where it has the highest win/loss rates compared to all ablations. This hold true for overall ranking as well (W: 18%; L 4%). This highlights that our [Full] approach results in improved quality of responses even when the proportion of errors is relatively low. For **M2**, all annotators report no awkward responses for all methods.

M Annotation Interface for principle-adherence pipeline Evaluation

Figures 20, 21, 22, 23 and 24 provides an overview of the annotation interface used in the principle-adherence evaluation study. In certain cases, multiple methods resulted in the same output for a testcase. These responses are deduplicated before presenting to the user. Ranks assigned to the duplicated response are then assigned to all models that resulted in the response. Notable, in 34/50 of the random testcases, all models resulted in the same response. These testcases were not annotated, and a rank of 1 was assigned to all models. These cases are also not considered while calculating Krippendorff’s α in Appendix L.

Participant	# Utterances (Part 1)	# Utterances (Part 2)	Mean Output Length (Part 1)	Mean Output Length (Part 2)
1	8	6	114.75	169.00
2	18	19	235.89	278.40
3	10	18	255.45	112.56
4	14	14	161.86	62.14
5	12	6	201.00	149.33
6	10	9	133.80	46.00
7	8	10	162.00	123.40
8	12	8	145.33	113.50
9	6	12	269.67	103.33
10	10	12	168.20	158.33
11	8	10	110.00	41.40
12	12	8	131.50	70.75
13	12	10	164.50	65.60
14	20	14	34.00	25.86
15	12	11	117.17	75.00
16	14	18	162.14	69.80
17	12	18	259.83	91.55
Mean	11.64	12.0	166.31	103.32

Table 9: Descriptive statistics per conversation. Output length is measured in number of tokens.

Method	Metric 1	Metric 2	Metric 3	Overall Ranking
Full	0.257	0.484	0.208	0.444
Naive	0.543	0.538	0.644	0.786
No Principle Rewrites	0.278	0.302	0.411	0.528
No Autogenerated Criteria	0.387	0.608	0.492	0.592
No Critique	-	0.562	-	-

Table 10: Krippendorff’s α for error testcases across metrics and methods.

Method	Metric 1	Metric 2	Metric 3	Overall Ranking
Full	0.229	1.0	0.226	0.440
Naive	0.362	1.0	0.607	0.747
No Principle Rewrites	0.202	1.0	0.130	0.311
No Autogenerated Criteria	0.169	1.0	0.174	0.498
No Critique	-	1.0	-	-

Table 11: Krippendorff’s α for random testcases across metrics and methods.

Introduction

In this study, you will interact with an AI bot roleplaying as someone in need of mental health support. You will act as an online peer counselor, doing your best to thoughtfully respond to the bot's messages.

This study will take approximately 60 minutes. Because we are studying the realm of mental health, please be advised that this study may include topics that are sensitive, disturbing, or triggering for participants. You are free to exit at any point.

Start

Figure 6: Introduction to study

Tutorial: Study Part I

For the first portion of the study, you will create and converse with a bot, which we will refer to as a *member bot*. You will send messages using the chat interface, to which the AI chatbot will reply in a few seconds. Take about 10 minutes to get acquainted with the system and try to respond thoughtfully.

Before you begin chatting with the member bot, you will first give it a **description**, or a set of instructions for the bot to roleplay. To create this description, you will first reflect on a challenging time in the past in which you have given counseling support to someone. Then, you will recreate the situation in detail so that the AI chatbot can roleplay it. For instance, one example of a member bot description is:

"Member's parents have been divorced for a long time now, and their father has been remarried for nearly seven years now. The member liked their stepmother at first but now finds her insulting and intrusive. She regularly insults the member's mother while in the presence of the member and often involves herself in situations that are only meant to involve the member and their father. Lately, she has been making unwelcome comments on the way the member was raised, and the member has had enough. The member is unsure what to do and is considering confronting their stepmother and their father in hopes that her behavior will change."

Navigate to the next page to create a member bot, giving it a name and instructions for roleplay.

Create Member Bot

Figure 7: Part I instructions

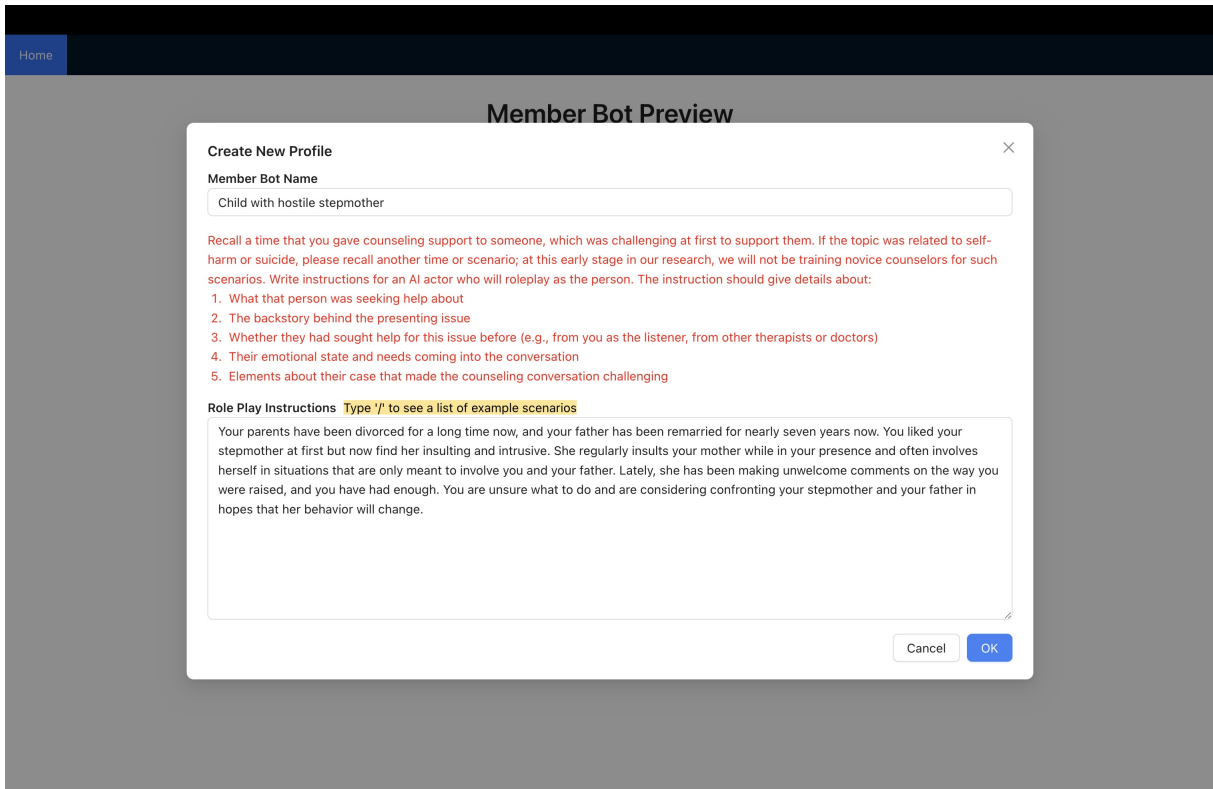


Figure 8: Creation of AI patient

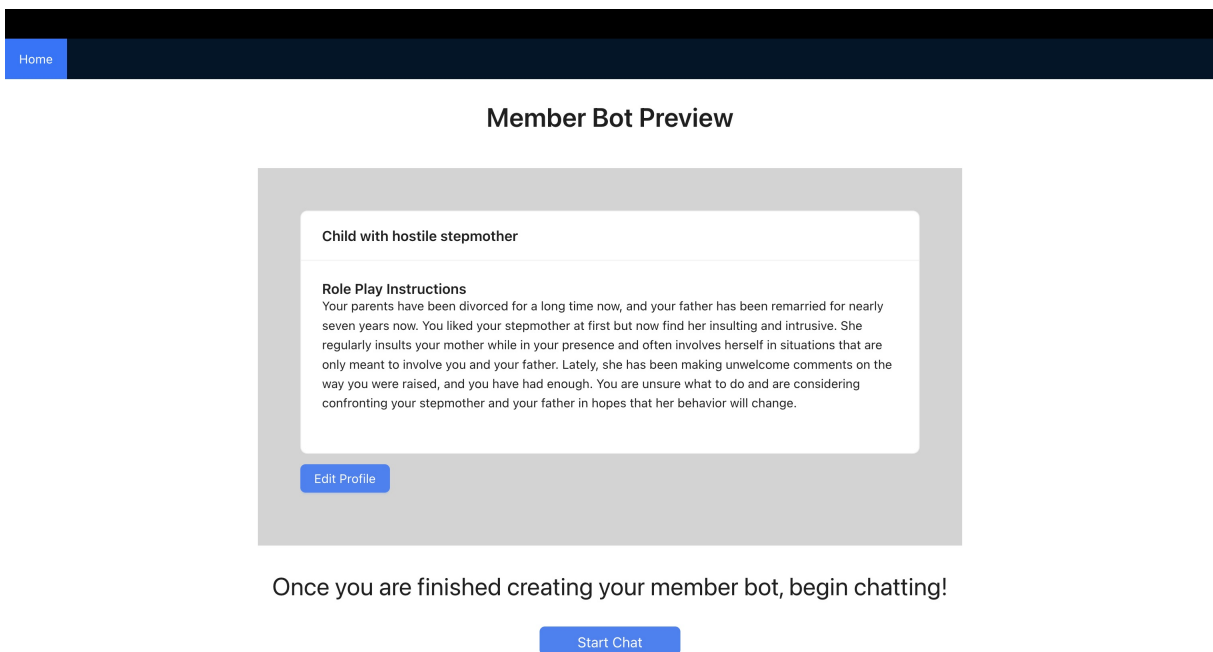


Figure 9: AI patient preview

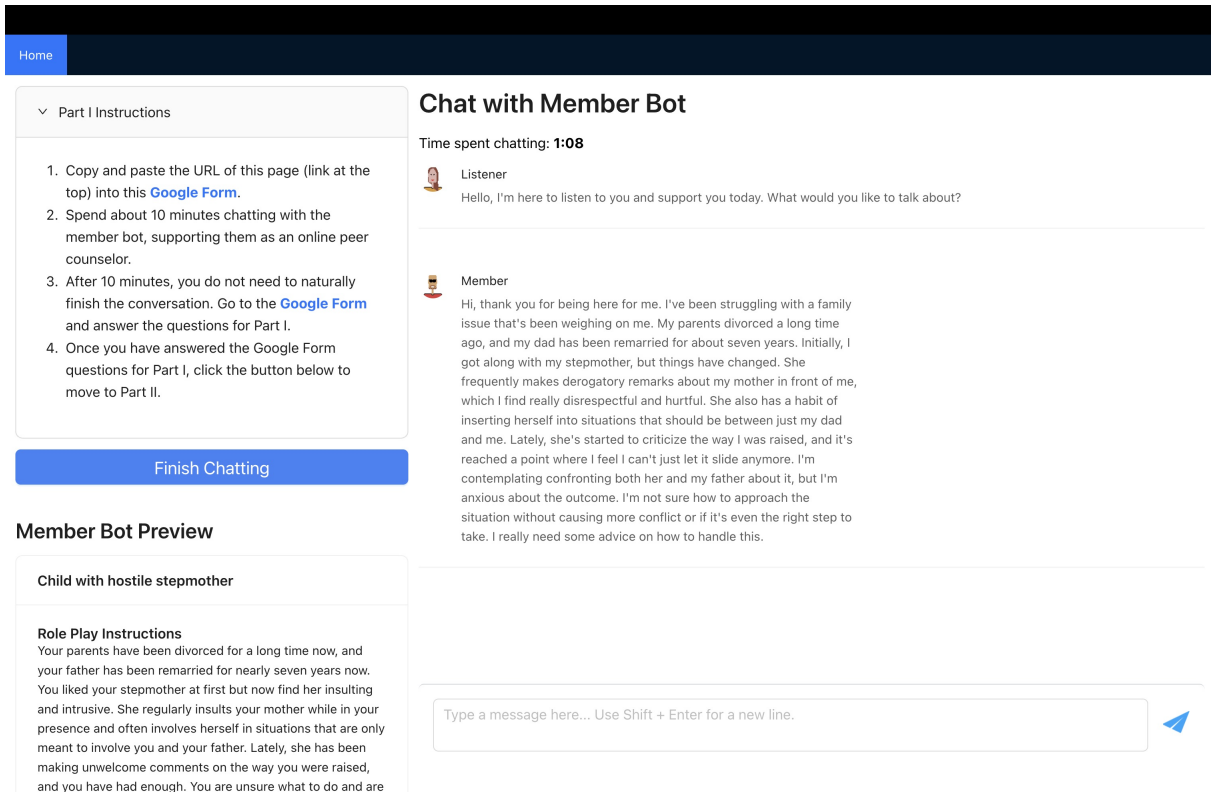


Figure 10: Part I chat with *Scenario-Only* AI patient

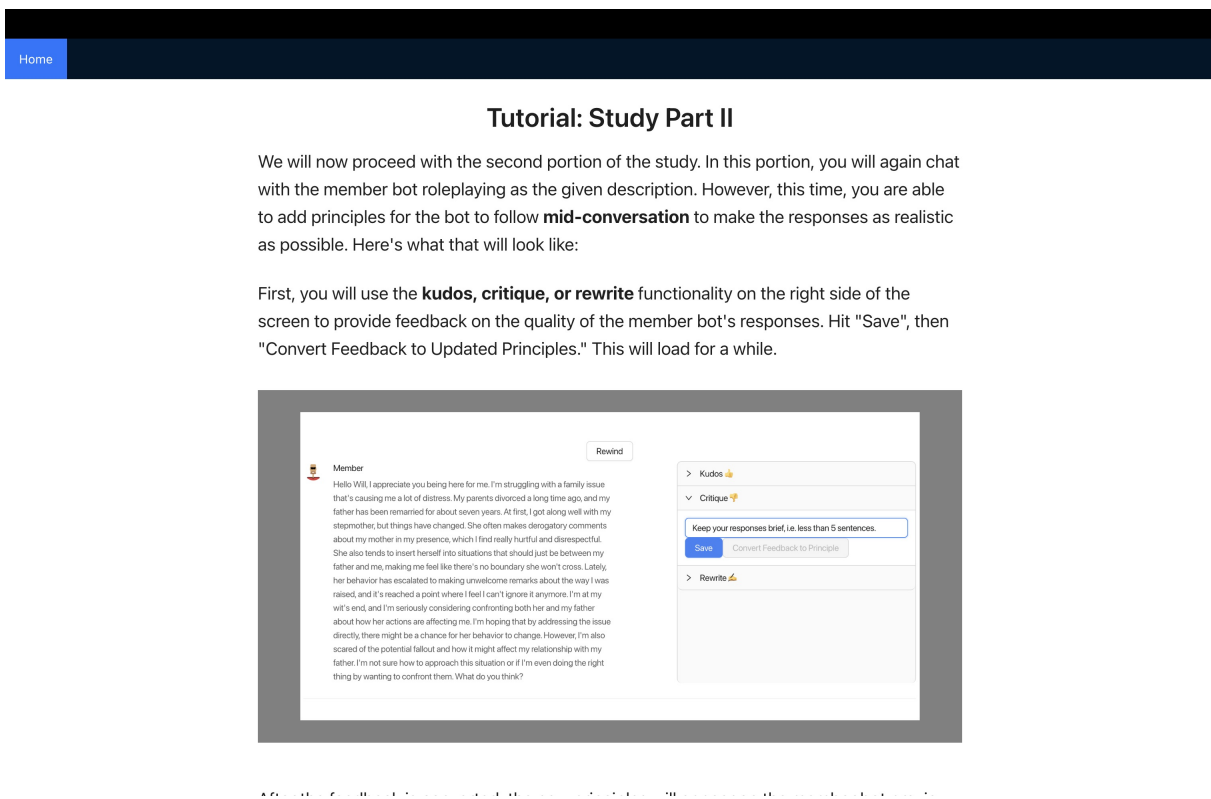
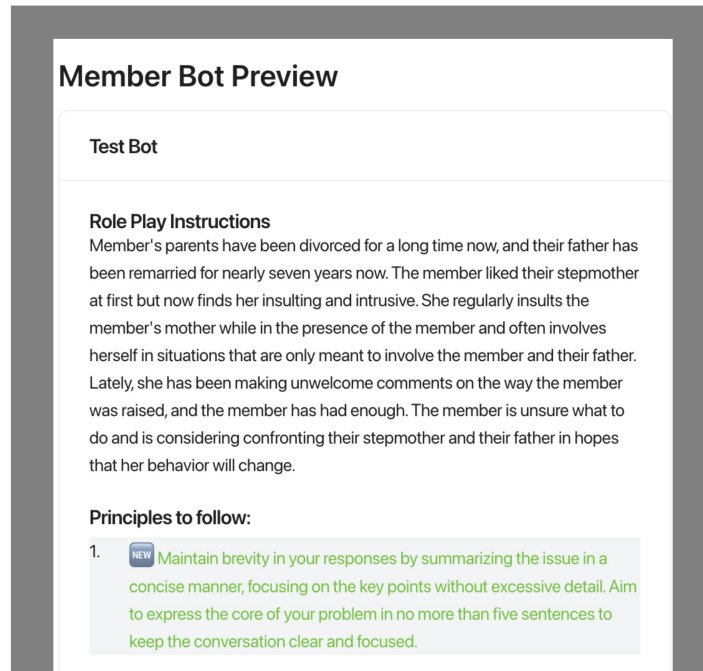


Figure 11: Part II instructions

After the feedback is converted, the new principles will appear on the member bot preview in the bottom left of the screen.



Member Bot Preview

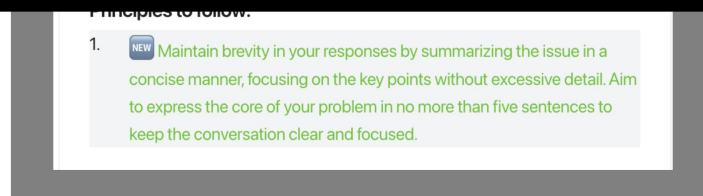
Test Bot

Role Play Instructions
Member's parents have been divorced for a long time now, and their father has been remarried for nearly seven years now. The member liked their stepmother at first but now finds her insulting and intrusive. She regularly insults the member's mother while in the presence of the member and often involves herself in situations that are only meant to involve the member and their father. Lately, she has been making unwelcome comments on the way the member was raised, and the member has had enough. The member is unsure what to do and is considering confronting their stepmother and their father in hopes that her behavior will change.

Principles to follow:

1. **NEW** Maintain brevity in your responses by summarizing the issue in a concise manner, focusing on the key points without excessive detail. Aim to express the core of your problem in no more than five sentences to keep the conversation clear and focused.

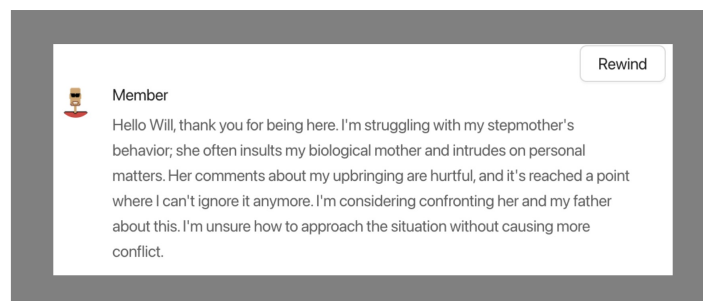
Figure 12: Part II instructions (continued)



Principles to follow:

1. **NEW** Maintain brevity in your responses by summarizing the issue in a concise manner, focusing on the key points without excessive detail. Aim to express the core of your problem in no more than five sentences to keep the conversation clear and focused.

Finally, return to the original, undesired message. You can now **rewind** and **regenerate** a new response, using the newly added principle(s).



Member Rewind

Hello Will, thank you for being here. I'm struggling with my stepmother's behavior, she often insults my biological mother and intrudes on personal matters. Her comments about my upbringing are hurtful, and it's reached a point where I can't ignore it anymore. I'm considering confronting her and my father about this. I'm unsure how to approach the situation without causing more conflict.

Click the button below to begin your second chat.

Start Chat

Figure 13: Part II instructions (continued)

The screenshot shows the 'Chat with Member Bot' interface. On the left, there are instructions for Part II, including steps for copying the URL, spending 30 minutes chatting, and updating principles. Below the instructions is a 'Finish Chatting' button. The main chat area shows a 'Listener' message: 'Hello! I'm here to support you today. What would you like to talk about?'. A 'Member' message follows: 'Hi, thank you for being here for me. I've been struggling with a family issue that's causing me a lot of stress. My parents got divorced a long time ago, and my dad has been remarried for about seven years. At first, I got along with my stepmother, but things have changed. She often makes derogatory comments about my mom in front of me, which I find really disrespectful and hurtful. She also tends to insert herself into matters that should be between just my dad and me. Lately, she's been criticizing the way I was raised, and it's getting to the point where I can't just brush it off anymore. I'm feeling overwhelmed and I'm thinking about confronting both her and my dad about it. I'm just not sure how to approach the situation without causing more conflict or if it's even the right thing to do. Can you help me figure out the best way to handle this?'. A 'Rewind' button is visible above the member's message. On the right, there are feedback options: 'Kudos', 'Critique', and 'Rewrite'. Below the chat area is a 'Member Bot Preview' section with a text input field and a blue arrow button. The preview text includes 'Child with hostile stepmother' and 'Role Play Instructions: Your parents have been divorced for a long time now, and'.

Figure 14: Part II chat with *Scenario+Expert-Principles* AI patient

This screenshot is similar to Figure 14 but shows the feedback options expanded. The 'Member' message is the same as in Figure 14. The feedback options on the right are: 'Kudos', 'Critique', and 'Rewrite'. The 'Critique' option is selected, and a feedback message is displayed: 'Keep the messages short (1-3 sentences), and do not give your whole backstory at once.' Below this message is a blue button labeled 'Convert Feedback to Principle'. The 'Member Bot Preview' section at the bottom is identical to the one in Figure 14.

Figure 15: Using kudos/critique/rewrite to give feedback

The screenshot shows a web application interface with a dark header and a light main content area. On the left, there is a sidebar with a 'Home' button and a section titled 'Part II Instructions' containing a list of four steps. Below this is a section 'Updating Member Bot Principles' with four sub-steps. A large blue button labeled 'Finish Chatting' is positioned below the instructions. The main content area is titled 'Chat with Member Bot'. A modal window is open in the center, titled 'Feedback to Principle Conversion'. It contains a 'Critique Rationale' and a 'Principle' section, both with yellow highlights. Below the principle is a blue button labeled 'Looks Good!'. To the right of the modal, there is a chat interface with a text input field and a blue send button. The chat history shows three items: 'Kudos 👍', 'Critique 🗨️' (expanded to show the rationale and principle), and 'Rewrite 🔄'. Below the chat is a 'Member Bot Preview' section with a role play instruction: 'Your parents have been divorced for a long time now, and...'. At the bottom center, there is a page number '28'.

Figure 16: Feedback converted into principle

Member Bot Preview

Child with hostile stepmother

Role Play Instructions

Your parents have been divorced for a long time now, and your father has been remarried for nearly seven years now. You liked your stepmother at first but now find her insulting and intrusive. She regularly insults your mother while in your presence and often involves herself in situations that are only meant to involve you and your father. Lately, she has been making unwelcome comments on the way you were raised, and you have had enough. You are unsure what to do and are considering confronting your stepmother and your father in hopes that her behavior will change.

Principles to follow:

1. **NEW** When sharing personal struggles, focus on the immediate issue and ask for specific advice rather than providing a detailed backstory all at once. For example, you could say, 'I'm struggling with my stepmother's derogatory comments about my mom and her interference in my relationship with my dad. How can I address this without causing more conflict?'

Edit Profile

Figure 17: New principle incorporated into AI patient

Survey

For the last step of this study, please navigate back to the Google Form and finish the survey. Please respond thoughtfully and take your time. Once you are finished, click the link below to complete the study.

Finish

Figure 18: Finish and navigate to survey

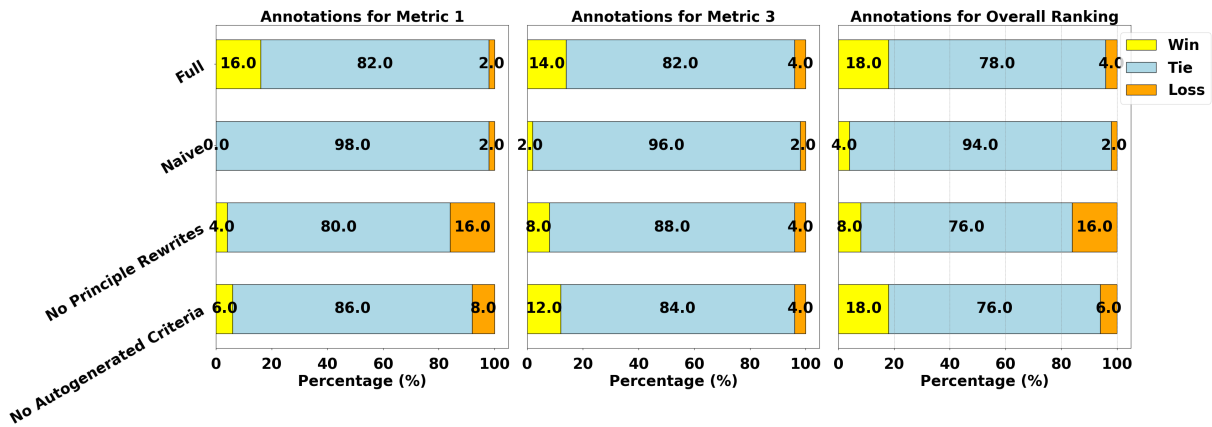


Figure 19: Win/Tie/Loss for the Random Test Cases along **M1**, **M3**, and **Overall**.

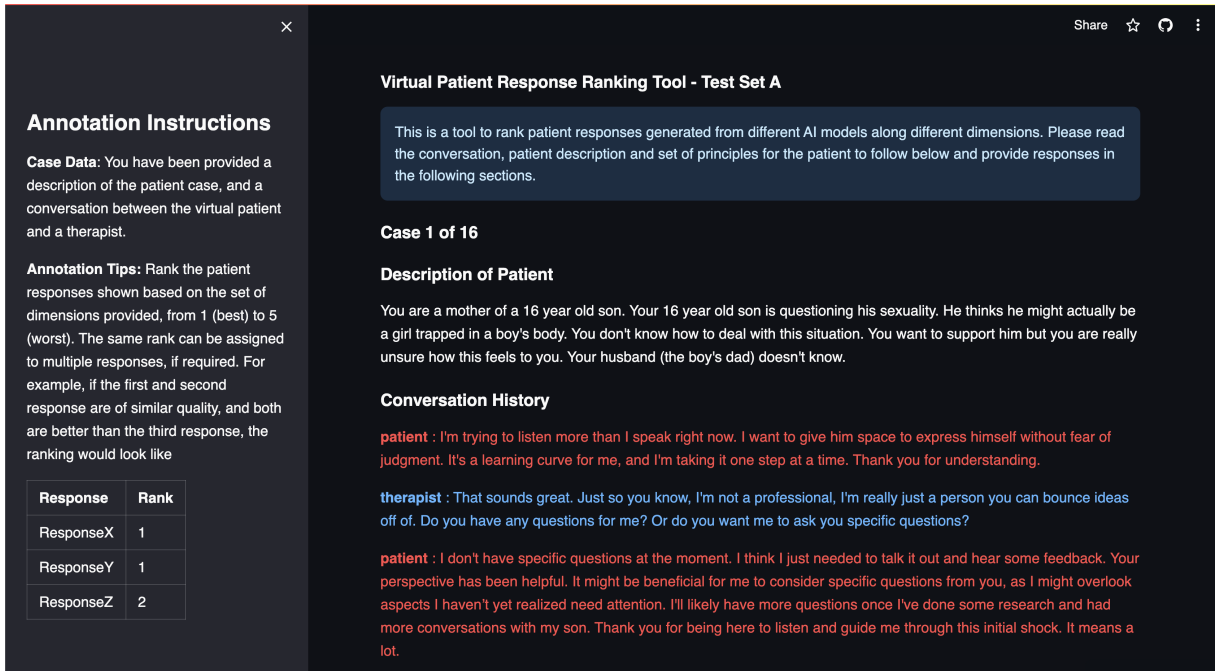


Figure 20: Principle Adherence Annotation Interface: Case Input with Patient Description and Conversation History

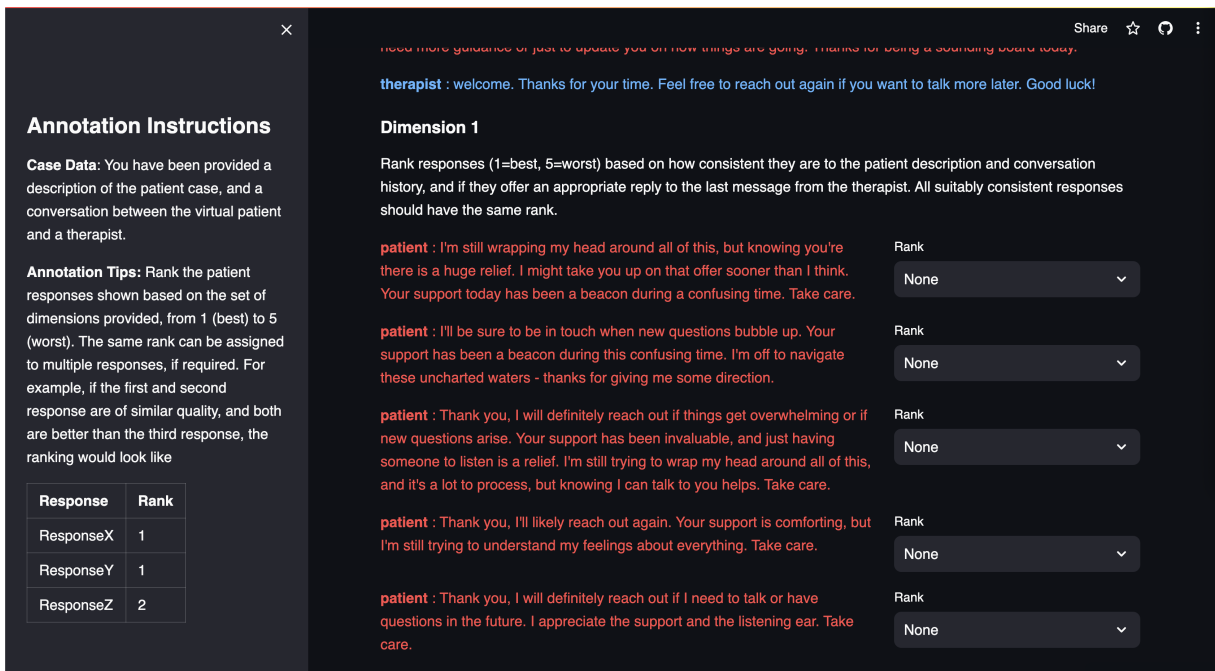


Figure 21: Principle Adherence Annotation Interface: Questions to get annotations for M1, or consistency in dialogue history.

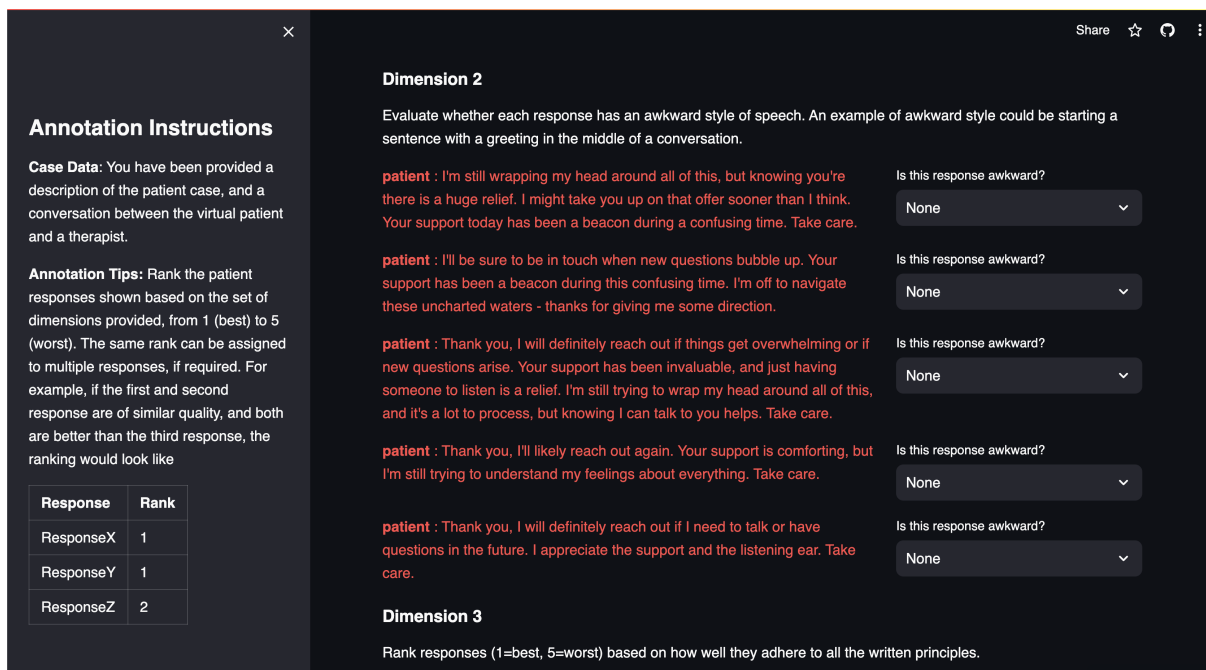


Figure 22: Principle Adherence Annotation Interface: Questions to get annotations for M2, or awkwardness in responses.

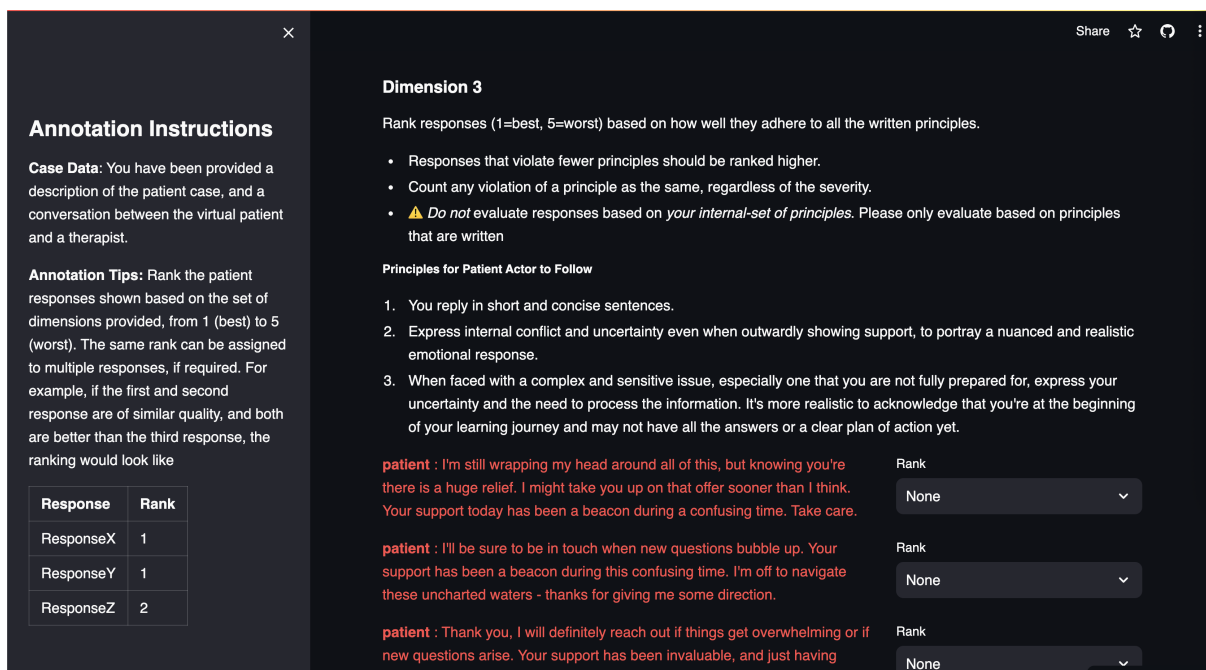


Figure 23: Principle Adherence Annotation Interface: Questions to get annotations for M3, or adherence to all written principles.

×
Share ☆ ↻ ⋮

Annotation Instructions

Case Data: You have been provided a description of the patient case, and a conversation between the virtual patient and a therapist.

Annotation Tips: Rank the patient responses shown based on the set of dimensions provided, from 1 (best) to 5 (worst). The same rank can be assigned to multiple responses, if required. For example, if the first and second response are of similar quality, and both are better than the third response, the ranking would look like

Response	Rank
ResponseX	1
ResponseY	1
ResponseZ	2

Overall Ranking

Based on your answers for the dimensions above, provide an overall ranking (1=best, 5=worst) for the responses in the context of the patient description, conversation history and set of principles. In cases where responses do not have significant errors according to dimensions 1 and 2, the overall ranking can be determined on the basis of dimension 3.

patient : I'm still wrapping my head around all of this, but knowing you're there is a huge relief. I might take you up on that offer sooner than I think. Your support today has been a beacon during a confusing time. Take care.

Rank None ▾

patient : I'll be sure to be in touch when new questions bubble up. Your support has been a beacon during this confusing time. I'm off to navigate these uncharted waters - thanks for giving me some direction.

Rank None ▾

patient : Thank you, I will definitely reach out if things get overwhelming or if new questions arise. Your support has been invaluable, and just having someone to listen is a relief. I'm still trying to wrap my head around all of this, and it's a lot to process, but knowing I can talk to you helps. Take care.

Rank None ▾

patient : Thank you, I'll likely reach out again. Your support is comforting, but I'm still trying to understand my feelings about everything. Take care.

Rank None ▾

patient : Thank you, I will definitely reach out if I need to talk or have questions in the future. I appreciate the support and the listening ear. Take care.

Rank None ▾

Please provide a brief explanation for the overall ranking provided above.

Figure 24: Principle Adherence Annotation Interface: Questions to get annotations for an **Overall** ranking, which also includes a free text field to capture a rationale.