PSI-Bench: Interpretable and Clinically Meaningful Evaluation of Depression Patient Simulators

Nguyen Khoi Hoang[1,2] **Shuhaib Mehri**[1] **Tse-An Hsu**[1] **Yi-Jyun Sun**[1]
Quynh Xuan Nguyen Truong[3] **Khoa D Doan**[2] **Dilek Hakkani-Tür**[1]

[1]University of Illinois Urbana-Champaign,
[2]VinUni-Illinois Smart Health Center,
[3]College of Health Sciences, VinUniversity

{nkhoang2, dilek}@illinois.edu

Abstract

Patient simulators are gaining traction in mental health training by providing scalable exposure to complex and sensitive patient interactions. Simulating depressed patients is challenging, as safety constraints and high patient variability complicate simulations and underscore the need for simulators that capture diverse and realistic patient behaviors. However, existing evaluations heavily rely on LLM-judges with poorly specified prompts and do not assess behavioral diversity. We introduce **PSI-Bench**[1], an automatic evaluation framework that provides interpretable, clinically meaningful diagnostics of depression patient simulator behavior across turn-, dialogue-, and population-level dimensions. Using PSI-Bench, we benchmark seven LLMs across two simulator frameworks and find that simulators produce overly long, lexically diverse responses, show reduced variability, and move through therapeutic stages and toward positive valence too quickly. We also show that the simulation framework has a larger impact on fidelity than the model scale. Results from a human study demonstrate that our benchmark is strongly aligned with judgments of mental health professionals. Our work reveals key limitations of current depression patient simulators and provides an interpretable, extensible benchmark to guide future simulator design and evaluation.

1 Introduction

Large Language Models (LLMs) are increasingly used to simulate human behavior in interactive settings where real-world practice is costly, sensitive, or difficult to scale (Park et al., 2023; Yao et al., 2024; Zhou et al., 2024). In mental health education, LLM-based patient simulators can generate clinically plausible responses and act as stand-ins for real patients, offering a practical training environment for novice clinicians (Liu et al., 2025; Wang et al., 2024; Louie et al., 2024). The fidelity of these simulators is critical, as divergence from human patient communication, such as in language use or conversational progression, can risk reinforcing misconceptions and undermining training effectiveness (Mehri et al., 2025; Abdulhai et al., 2025; Naous et al., 2026). Therefore, it is essential to evaluate simulator fidelity with human patient behavior. This work focuses on a particularly challenging domain of mental health: depression. Simulating depression is difficult as LLMs' safety training limits their ability to mimic depressive speech patterns (Liu et al., 2025) and as depression varies widely between individuals (Fried and Nesse, 2015).

Current automatic evaluations of simulator fidelity rely on LLM-judges to assess how well generated responses align with the target patient profiles (Liu et al., 2025; Wang et al., 2024). They score responses on a 5-point Likert scale (Likert, 1932) using generic prompts, entrusting the LLM-judge to infer nuanced clinical distinctions without explicit grounding in psychological theory or human patient data. This setting may not be reliable, as Wang et al. (2024) reports opposite trends between LLM-judge fidelity scores and expert ratings. Moreover, it lacks the interpretability needed to reveal *where* a simulator diverges from human patient behavior, offering little guidance for improvement. Additionally, existing evaluation approaches (Sabour et al., 2026; Wang et al., 2024; Liu et al., 2025) report an average score without examining whether the simulators reflect the distributional diversity of patient populations. Patients can exhibit varying linguistic and interactional patterns (Ruiz-Yu et al., 2025), especially among depressed patients (Fried and Nesse, 2015). Simulators that fail to reflect such diversity may introduce bias in clinician training (Schwieters et al., 2023).

Recognizing these limitations, we propose Patient-SImulator-Bench (PSI-BENCH), a framework for interpretable, clinically meaningful eval-

[1]We release the code and simulated conversation datasets at: https://github.com/nguyenkhoihoang/PSI-Bench

Proceedings of the 1st Workshop on Multimodal, Multilingual, and Multicultural Mental Health and Psychotherapy (MultiPsyche), pages 118–144
November 9, 2026 ©2026 Association for Computational Linguistics

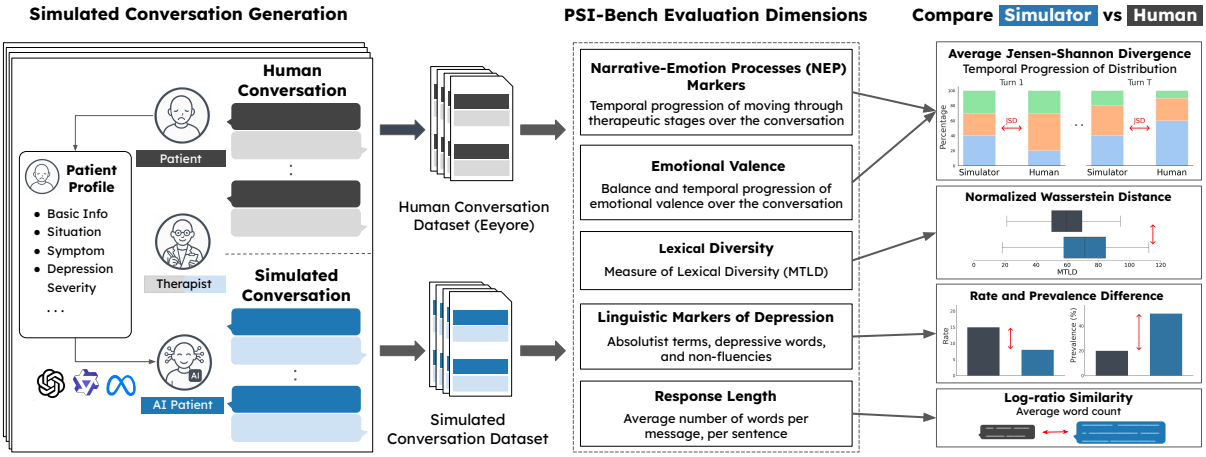


Figure 1: Overview of PSI-Bench. From the Eeyore dataset (Liu et al., 2025) of depression-related human patient conversations, we condition the patient simulator on the patient profiles derived from these conversations to generate a simulated conversation dataset. Both the human and simulated datasets are then evaluated independently using the PSI-Bench dimensions, and the resulting statistics are compared.

uation of depression simulators. PSI-Bench comprises five dimensions supported by established psychological and psycholinguistic research (Rude et al., 2004; Trifu et al., 2024; O'Dea et al., 2021; Tølbøll, 2019; Smirnova et al., 2018; Beck, 1967), ensuring that the evaluation meaningfully captures the characteristics of depression behavior. Given a set of reference human conversations and patient profiles derived from them, we evaluate generated and human responses along these dimensions to quantify simulator alignment with human patient behavior. Due to privacy and ethical constraints on accessing real clinical data, we apply PSI-Bench to Eeyore (Liu et al., 2025), an expert-curated English corpus of crowd-sourced and acted depression-related conversations paired with extracted patient profiles. Such proxy interactions are well-established in mental health education (Dawood et al., 2024), and the framework can be applied to authentic conversation datasets as they become available. To assess the reliability and diagnostic value of PSI-Bench, we use PSI-Bench to investigate three research questions:

- **RQ1:** How do responses generated by current depression patient simulators differ from human reference across turn-level, dialogue-level, and population-level?
- **RQ2:** How do different LLMs and simulation strategies affect simulator fidelity?
- **RQ3:** Do the findings from answering RQ1 and RQ2 using PSI-Bench align with judgments of mental health professionals?

We evaluate seven LLMs across two simulation frameworks (Wang et al., 2024; Louie et al., 2024), and reveal key limitations not captured by previous automatic evaluation approaches. At the population level, simulators show less variability and diverge from human reference in dialogue progression: they move through therapeutic stages more rapidly and follow a uniform negative-to-positive valence trajectory. They also produce overly long, lexically diverse responses at the message level, and exhibit a more spread out presence of linguistic markers of depression at the dialogue level.

We aggregate scores across all dimensions into a unified alignment score to rank the simulators (Table 1), while retaining interpretable diagnostics of each evaluation dimension. We find that the simulation framework can have a larger impact than LLM capability, and smaller models can be closer to human patient communication than larger ones. Finally, we validate our benchmark via a human study with 20 mental health professionals. The results show near-perfect agreement between their pairwise preferences and PSI-Bench's ordering ($\kappa = 0.82$), supporting PSI-Bench as a reliable evaluation framework. Our main contributions are:

- We introduce PSI-Bench, an interpretable, clinically meaningful benchmark for evaluating patient simulators by alignment with human reference, showing strong agreement with judgments of mental health professionals.
- Using PSI-Bench, we highlight previously underexplored simulator divergences from human patient behavior, and quantify the impact of simulation framework, scale, and model family on simulator fidelity.

2 Related Work

2.1 LLM-based Simulators

User Simulators for Dialogue Systems User simulators have been studied since the mid-1990s to address the cost and difficulty of collecting real conversational data, particularly in specialized domains (Pietquin and Hastie, 2013). While traditional rule-based, statistical, and early neural approaches struggled to capture natural human behavior, LLMs offer powerful language capabilities to overcome these limitations (Naveed et al., 2025). Existing work mainly explores two applications of LLM-based simulators. First, they are used to *train and evaluate AI systems* by generating dialogue data (Finch and Choi, 2024; Bao et al., 2023) or as interactive environments for online reinforcement learning (Carta et al., 2023). Second, they are used as *training tools for human learners* in domains where practicing with real individuals is costly or sensitive, such as nurse communication training (Lee et al., 2025) and counseling (Na et al., 2025).

LLM-based Patient Simulators in Mental Health Within the counseling domain, several works (Chen et al., 2023; Lee et al., 2024) focus on the first application – generating counseling datasets to improve counselor LLMs, where a patient simulator's design and evaluation are less prioritized. In contrast, our evaluation framework targets the second application – simulating patients for training novice counselors, where the central emphasis is on designing simulated patients with realistic behaviors (Liu et al., 2025; Wang et al., 2024; Louie et al., 2024).

2.2 Evaluation of Patient Simulators

Evaluating User Simulators Before the advent of LLMs, Pietquin and Hastie (2013) grouped the evaluation of user simulators into two groups: *turn-level* metrics that evaluate individual responses (e.g., precision and recall) and *dialogue-level* metrics that evaluate general interaction (e.g., task completion). Though fast and reproducible, these traditional metrics rely on surface lexical overlap and miss deeper nuances (Gu et al., 2025). Along with the rise of LLM-based simulators, LLM-based evaluation methods emerged with their strong reasoning and language capabilities, enabling more context-aware assessment but also introducing new systematic bias (Gu et al., 2025; Gallegos et al., 2024). Evaluation can also be framed as *intrinsic* versus *extrinsic*: intrinsic methods assess simulator outputs directly (e.g., realism and diversity), while extrinsic methods measure downstream task impact. Naous et al. (2026) show that extrinsic evaluation alone can be misleading as over-cooperative LLM users may inflate system performance, motivating greater emphasis on intrinsic methods.

Evaluating LLM-based Patient Simulators in Counseling Simulation systems aimed at training or benchmarking counselor LLMs (Lee et al., 2024; Wang et al., 2025) are often automatically evaluated using extrinsic metrics of counseling effectiveness. However, consistent with the findings of Naous et al. (2026), such simulated patients can be implicitly cooperative and demonstrate rapid improvement (Abdulhai et al., 2025), thus increasing extrinsic performance scores.

Patient simulators for training novice counselors are typically evaluated for alignment, realism, and training utility (Liu et al., 2025; Wang et al., 2024; Louie et al., 2024; Sabour et al., 2026). Expert human evaluation assess profile alignment, behavioral fidelity, and pedagogical effectiveness (Wang et al., 2024). While it is often impractical to use expert evaluation alone as it is costly and not scalable, current automatic evaluation pipelines which rely on LLM-Judge Likert ratings are reported to be unreliable or even contradict expert judgments (Wang et al., 2024). Additionally, they typically report an average score without examining broader behavioral distributions reflecting real-world clinical variability. Our evaluation framework addresses these limitations.

3 PSI-Bench

Figure 1 provides an overview of PSI-Bench. We use the reference conversation dataset's pre-extracted patient profiles by default; if a simulator provides its own profile-extraction method, we instead derive the patient representations from the reference dataset using that method. We generate simulated conversations by conditioning patient simulators on the patient profiles and pairing them with an LLM therapist. Both human reference and synthetic datasets are then evaluated independently using PSI-Bench dimensions, and the resulting statistics are compared.

Our reference dataset combines three dyadic conversation datasets from Eeyore (Liu et al., 2025), curated from public depression-related dialogue sources by extracting depressive traits to guide

data filtering and construct structured psychological profiles paired with each conversation (see Appendix A.1). While the conversations contain crowd-sourced and acted interactions, such proxy interactions are commonly used in mental health education via standardized patients due to privacy constraints (Dawood et al., 2024), and instruction-tuning on these conversations improved simulator realism (Liu et al., 2025). PSI-Bench can be readily extended to datasets containing authentic clinical conversations as such resources become available.

We ground our assessment in behaviors observed in depression-related communication and analyze simulator behavior across five interpretable dimensions. **Section 3.1** introduces and independently applies these dimensions to reference and synthetic conversations. **Section 3.2** describes how simulators' scores are compared relative to human patient scores at multiple levels: *turn-level* (e.g., response length), *conversation-level* (e.g., progression and valence change), and *population-level* (e.g., diversity across profiles), capturing realistic patterns and variation. Each dimension is aggregated into normalized scalar scores for comparison, ranking, and correlation with human evaluation.

3.1 Evaluation Dimensions

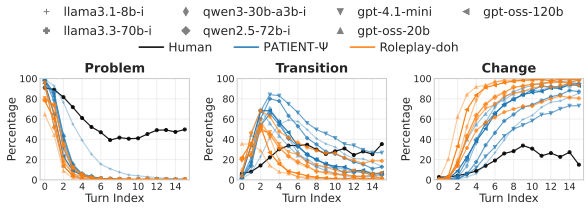
Narrative-Emotion Processes (NEP) Markers are standardized categories rated by observers that capture how patients manifest narrative content and emotional experience during psychotherapy (Angus et al., 2017). Because the evolving pattern of these markers over a session reflects therapeutic progress, they provide a basis for assessing whether simulated patients replicate the change dynamics observed in real clinical conversations. The markers are grouped into three categories: *Problem*: repetitive, "stuck" storylines with limited reflection; *Transition*: more coherent and reflective storytelling as clients begin to examine their thoughts and emotions; *Change*: clear shifts in emotional response and thought patterns, where clients may show a new understanding of key events in their life. To track these patterns, we use `gpt-oss-120b` (OpenAI, 2025) to classify each patient turn in conversations as Problem, Transition, Change, or Filler (neutral procedural replies without meaningful therapeutic content), using explicit definitions and decision criteria (Appendix A.3).

Emotional Valence Tølbøll (2019); Rude et al. (2004) show that depression correlates with increased negative and decreased positive emotion words. We therefore investigate whether simulated and human patients differ in how their emotional valence manifests and evolves over the course of the conversation. We use `gpt-oss-120b` to classify each patient message as Positive, Negative, or Neutral (Appendix A.4). We validate `gpt-oss-120b`'s NEP and valence classifications against annotations from 20 mental health professionals and find high agreement that supports its classification reliability (Section 6).

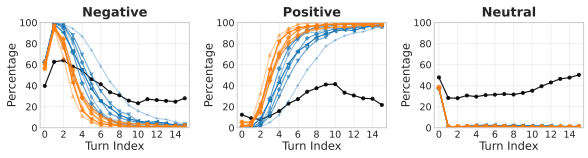
Lexical Diversity Smirnova et al. (2018) showed that patients with depression often exhibit repetition in words or ideas. We evaluate whether simulators can mimic this behavior despite LLMs' tendency to generate highly diverse text (Li et al., 2025). We also investigate how lexical diversity varies across human patients and whether the same LLM conditioned on different profiles can mirror such variation. We compute conversation-level lexical diversity by concatenating all patient messages in a conversation and applying the Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis, 2010). MTLD is defined as the mean length of word sequences required to reach a fixed ratio (0.72) of unique words to total words (Appendix A.5). Additionally, we concatenate all patient messages across profiles into a single corpus and compute its MTLD. A lower corpus-level MTLD than the average per-conversation MTLD would suggest some degree of shared vocabulary across patients.

Response Length Patients with depression tend to produce shorter utterances and more limited content in conversation (Trifu et al., 2024; Smirnova et al., 2018), while LLMs tend to be more verbose (Li et al., 2025). We assess response length by computing average number of words per message and average number of words per sentence.

Linguistic Markers of Depression Psychological research identifies consistent linguistic patterns in depressed patients that correlate with increased use of: (1) *absolutist terms*: words that may reflect absolutist thinking underlying many cognitive distortions (e.g. "always", "never") (Al-Mosaiwi and Johnstone, 2018), (2) *depressive words* associated with low mood and hopelessness (Li et al., 2020), and (3) *non-fluencies*: hesitations and fillers that may signal cognitive or emotional strain (O'Dea et al., 2021). Table 9 details the complete lexicon used. These markers enable a structured compar-



(a) Population-level distribution of narrative-emotion markers. Simulators show a more rapid progress than human reference.



(b) Population-level distribution of emotional valence. Human reference shows more neutral and varied valence, while simulators show a consistent negative-to-positive progression.

Figure 2: Population-level progression of patient responses across turns. Each point represents the percentage of messages assigned to a category at a given turn.

ison of depressive linguistic tendencies between humans and simulators. We analyze the proportion of messages containing at least one marker (marker prevalence) and the rate of marker occurrences per 1,000 tokens (marker rate).

3.2 Comparative Evaluation

We compare simulator scores with human reference statistics to quantify alignment with human patient behavior. All metrics are converted to 0–100 similarity scale, where higher values indicate higher alignment. We convert distance or divergence measures d to similarity scores: $100 - d$. We denote simulator statistics by ψ and reference by r.

Divergence of progression dynamics For NEP markers and valence, we compare the temporal progression of label distributions. To ensure comparability, we restrict the analysis to the first $T = 16$ turns for both human and synthetic conversations, yielding aligned turn-level distributions. For each turn i, we compute a population-level distribution on labels by aggregating all patient messages at that turn across profiles. This produces a sequence of T turn-level distributions $(\psi_i)_{i=1}^{T}$ for a simulator and $(r_i)_{i=1}^{T}$ for the reference patient population, where each ψ_i and r_i is a distribution on labels at turn i. This formulation enables us to track how the proportions of the labels evolve over the conversation. To quantify the divergence in progression dynamics between a simulator and the reference human population, we compute the symmetric Jensen-Shannon divergence (JSD) (Menéndez et al., 1997) at each

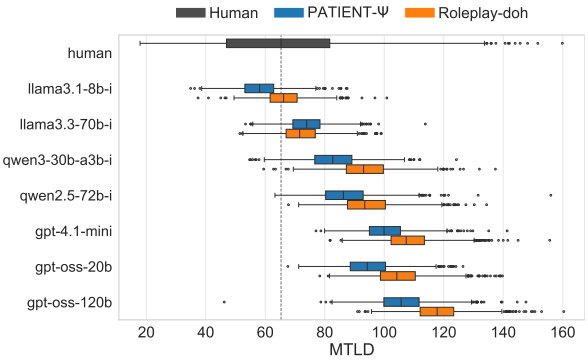


Figure 3: Boxplots of lexical diversity of human and simulated patient conversations. Simulators consistently show higher and more uniform lexical diversity than human patients.

turn and average across all turns:

$$\mathrm{d} = \frac{1}{T}\sum_{i=1}^{T} \mathrm{JSD}(\psi_i \,\|\, r_i) \tag{1}$$

Normalized Wasserstein distance For lexical diversity, we compare the distribution of per-conversation MTLD scores between simulator and reference using Wasserstein distance (Villani, 2009), which captures difference in both distribution shape and spread. We convert this distance into a similarity measure and apply min–max normalization.

Log-ratio similarity score As simulated patients are more verbose than human patients by large factors, we use log scale to account for this disparity. We measure response length alignment using a log-ratio similarity over average words per message and per sentence. For each average word count c, we compute:

$$s(c_\psi) = 100 \cdot \exp\left(-\left|\log \frac{c_\psi}{c_\mathrm{r}}\right|\right). \tag{2}$$

The final response length similarity score is the average of the two similarity scores.

Difference from human statistics For linguistic markers of depression, we measure the deviation from human patient statistics using the relative (percentage) difference for the marker rate and the absolute difference for the marker prevalence, as it is already a bounded percentage. The final distance d is the average of the two.

4 Experimental Setup

We evaluate two main counsellor-training patient simulation frameworks, instantiated with one

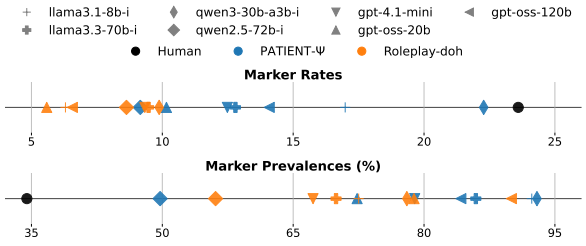


Figure 4: Dot plots of depression marker rates (markers count per 1,000 tokens) and prevalences (% messages containing marker). Simulators have lower marker rates but higher prevalences than human reference.

proprietary and six open-source LLMs spanning three model families (Table 8), for a total of 14 framework-LLM pairings. We use the same counsellor LLM `gpt-oss-20b` (OpenAI, 2025) across all settings. Generation is capped at 16 turns or the number of patient turns in the corresponding reference conversation, whichever is smaller. We repeat the benchmark with a second full generation run under identical settings, and another with `PsyCoPref-Llama3-8B` (Zhang et al., 2026) as the counsellor LLM. The **findings in Section 5 are consistent across all three runs**.

PATIENT-Ψ (Wang et al., 2024) is a cognitive behavioral therapy–based patient simulation framework that constructs diverse patient *cognitive models* – a structured and interconnected representation of thoughts, beliefs, emotions, and behaviors.

Roleplay-doh (Louie et al., 2024) is a human-LLM collaborative framework that enables experts to customize patient simulators by converting qualitative feedback into natural language "principles" governing behavior. It utilizes a specialized principle-adherence pipeline to verify and self-refine responses to ensure alignment with these expert-defined rules and the underlying patient persona.

5 Results

5.1 RQ1: Divergence from Human Reference

Comparing human and synthetic conversations across our evaluation dimensions reveals clear divergences in simulator behavior from the reference human patients (More results in Appendix C):

Early resolution of emotional narratives Figure 2a displays the progression of population-level distributions of NEP markers across dialogue turns. This figure shows that simulated patients progress through narrative–emotion stages much faster than human. Although all populations mostly start the conversation in the *problem* stage, human patients transition gradually, with over 40% remaining in this stage throughout the first 16 turns. In contrast, simulators quickly shift away from *problem* by around turn 3, with rapid percentage increases in *transition* and *change*. This suggests earlier narrative resolution and greater emotional awareness than human patients, who typically require longer interaction to make therapeutic progress.

Uniform, rapid negative-to-positive progression Figure 2b displays the progression of population-level valence distributions across dialogue turns. Negative valence dominates simulator messages in early turns but declines rapidly as the proportion of positive messages increases by the fourth turn. This rapid shift from negative to positive affect is consistent with the rapid progression observed in Figure 2a. In contrast, human valence is more varied and stable: negative valence appears early but does not dominate and declines gradually, while the percentage of positive messages rises modestly before declining as neutral messages increase.

Lack of conversational neutrality & spontaneity Human reference conversations frequently include filler messages and neutral responses, with roughly 30–50% of profiles having neutral valence at any given turn. In contrast, simulator outputs almost always map to a NEP marker and a clear valence, with barely any filler or neutral content beyond the first turn. This suggests that simulators lack the human conversational naturalness that often includes more neutral, spontaneous behavior, instead always producing task-focused and emotionally explicit responses. As this tendency exists for all simulator settings, in Section 5.2 we re-normalize distributions by excluding *filler* and *neutral* messages.

Higher lexical diversity, reduced variability Figure 3 shows that human patients exhibit lower median and more widely distributed MTLD scores, reflecting substantial individual variability. In contrast, almost all simulated patients produce higher and tightly ranged MTLD values, indicating uniformly diverse output that fails to capture the lower diversity and greater variability of human patient distribution. Table 3 in Appendix compares mean per-conversation diversity to corpus-level diversity. Interestingly, only human data shows a considerable drop at the corpus level, while simulated patients show similar scores for both. This may suggest that human patients have some lexical overlap

Rank	Setting	NEP ↑	Valence ↑	Div. ↑	Length ↑	Markers ↑	Overall ↑
1	PS-llama3.1-8b-i	**70.70**	**80.83**	72.17	28.24	57.05	61.80
2	PS-qwen3-30b-a3b-i	54.78	68.63	63.17	**43.89**	**67.97**	59.69
3	PS-qwen2.5-72b-i	56.78	67.95	57.25	<u>42.88</u>	<u>61.76</u>	57.33
4	RD-llama3.3-70b-i	57.01	66.98	<u>72.86</u>	29.41	52.35	55.72
5	PS-llama3.3-70b-i	58.93	73.04	70.57	24.18	51.37	55.62
6	RD-llama3.1-8b-i	52.37	65.27	**73.43**	25.24	44.33	52.13
7	RD-qwen2.5-72b-i	52.28	66.51	44.70	37.60	57.47	51.71
8	PS-gpt-4.1-mini	<u>59.68</u>	<u>75.80</u>	33.07	33.84	54.22	51.32
9	PS-gpt-oss-20b	52.80	68.30	43.74	28.76	52.59	49.24
10	RD-qwen3-30b-a3b-i	46.66	61.81	45.98	35.21	49.15	47.76
11	PS-gpt-oss-120b	55.40	74.17	22.66	26.00	54.97	46.64
12	RD-gpt-4.1-mini	54.24	69.29	18.98	30.95	53.37	45.36
13	RD-gpt-oss-20b	43.42	62.84	24.97	22.72	39.64	38.72
14	RD-gpt-oss-120b	46.11	63.52	0.00	20.77	36.14	33.31

Table 1: Benchmark results across all dimensions and average overall score. Higher scores indicate higher alignment with the human reference. *PS–* : PATIENT-Ψ, *RD–* : Roleplay-doh, *Div.*: lexical diversity. Rankings are based on overall score (descending). **Bold** and <u>underline</u> denote the best and second-best results, respectively.

between patients while simulated patients do not.

Higher verbosity Human patient messages are concise (averaging 18 words/message, 9 words/sentence), while simulated patients are substantially more verbose (64 – 319 words/message, 15 – 28 words/sentence).

Higher prevalence but lower density of depression markers Combined with divergence in verbosity, Figure 4 highlights differences in how linguistic markers of depression are expressed. Simulated patients consistently lie to the left of human data in marker rates but to the right in marker prevalence, indicating that markers appear in more messages but with lower token-level density. This pattern suggests that simulators repeatedly signal depression cues across many messages, but distribute them sparsely within longer utterances. In contrast, human patients express markers in fewer messages but at higher token-normalized rates, indicating that when markers do appear, they are more densely concentrated within shorter utterances.

5.2 RQ2: Impact of Simulation Methods and LLM Choice on Fidelity

To facilitate comparison across simulation settings, we aggregate all dimension similarities into a single overall alignment score in Table 1. Adjacent settings often have close scores (median 1.6 points gap). A repeated generation run under identical settings reproduces the ranking exactly (Table 6), while changing the counsellor LLM shifts closely scored settings by at most two positions (Table 7). We therefore focus on broader patterns that remain consistent across all three benchmark runs, examining the effects of simulation framework, model scale, and model family below.

Simulation framework has a higher impact than LLM. Results show that the top three configurations are all PATIENT-Ψ (PS) variants, and PS consistently outperforms Roleplay-doh (RD) under the same LLM backbone, except `llama-3.3-70b-i` where the two variants' scores are nearly tied. This suggests that framework design may play a bigger role than model choice, motivating greater attention to improving simulator design in future work.

Larger models do not reliably improve simulation quality. In several cases, smaller models outperform larger counterparts under the same framework. This trend is especially clear when comparing models within the same family; for instance, `gpt-oss-20b` outperforms `gpt-oss-120b` for both PS and RD frameworks. We hypothesize that larger models, due to their stronger capability, tend to produce responses that are overly coherent and self-aware, which may resemble therapeutic guidance rather than authentic patient expression.

Model family influences performance. Across model families, Llama-based models generally achieve higher alignment scores, followed by Qwen models, while GPT-based models tend to score lower overall.

6 Human Study

6.1 Study Design

To assess the reliability of PSI-Bench and answer RQ3, we conducted a human evaluation with 20 mental health professionals recruited via Prolific. Annotators were divided into four groups of five, each evaluating conversations from a different set

	Human–Model		Human–Human
Task	**% Agreement**	**Cohen's κ**	**Fleiss' κ**
Pairwise	91.67	0.8210	0.4328
NEP	80.06	0.6958	0.4997
Valence	85.33	0.725	0.5712

Table 2: Human–model agreement and inter-rater agreement across evaluation tasks.

of three patient profiles. For each profile, annotators completed two tasks: (1) *message-level classification* of emotional valence[2] and narrative-emotion process, to assess the reliability of our LLM-based classifications; (2) *pairwise comparison* of simulated conversation snippets, to assess agreement with our benchmark rankings.

To assess whether PSI-Bench's broader ordering aligns with mental health professionals' judgments, we randomly sample conversations from simulator pairs separated by at least four ranks. We use rank-separated pairs because adjacent settings are often closely scored and can exchange positions under design choices such as the counsellor LLM (Table 7); thus, the benchmark is not intended to distinguish near-tied systems by exact rank. Given two simulated conversation snippets of the same patient profile, annotators select which better reflects how a real patient would communicate, and alignment is measured by whether annotators consistently prefer outputs from higher-ranked simulators. For both tasks, we use majority voting within each group to determine the final human annotation. More details are provided in Appendix B.

6.2 RQ3: Alignment with Annotators

Table 2 reports agreement between mental health professionals' judgments and our benchmark rankings and LLM-based classifications. We report: (1) model-human agreement using percentage agreement and Cohen's κ, which measures agreement between two raters while accounting for chance, and (2) agreement among human annotators using Fleiss' κ, widely used to assess inter-rater reliability among multiple annotators. Following the interpretation of Landis and Koch (1977), inter-annotator agreement is moderate across tasks (Fleiss' κ between 0.43 and 0.57). However, agreements between our methods and annotators' majority labels are high: pairwise preference shows almost perfect alignment ($\kappa = 0.82$), while NEP marker and valence classification show substantial agreement ($\kappa = 0.7$ and 0.73, respectively). These results confirm that PSI-Bench provides reliable, annotator-aligned evaluations.

Annotators' qualitative feedback corroborates Section 5.1's findings on divergences from human patient behavior. Less realistic conversations were described as overly polished and cognitively organized. They resemble a "coherent story" rather than spontaneous distress, often featuring sophisticated vocabulary and lacking hesitation or disfluency. Annotators also noted that simulators were excessively self-reflective and articulate, which was rare in early interactions where real patients tend to struggle with articulation. Additionally, simulated responses were often prematurely solution-oriented, proposing coping strategies or rational interpretations, unlike real patients who may initially feel overwhelmed or cognitively blocked.

While annotators more often justified their choices by critiquing the less realistic option, some identified reasons certain conversations appeared more realistic. These were described as shorter, less certain, and more fragmented, using simpler language and exhibiting hesitation, self-correction, and slightly disjointed thought patterns. Notably, these conversations largely came from our top-ranked settings. In one instance, an annotator remarked that a conversation from our top-1 setting "seems most human and the realest of all the conversations in this block" (Table 12). These feedback further provide qualitative support for PSI-Bench's alignment with annotators' judgments.

7 Conclusion

We introduce PSI-Bench, an interpretable, clinically meaningful evaluation framework that decomposes simulator behavior into psychologically meaningful dimensions and evaluates alignment with reference human patient conversations. We reveal underexplored divergences between simulated and human patients, showing that simulators are less varied, transition through therapeutic stages and become positive more quickly, and produce overly verbose, lexically rich responses compared to human patients. We also find that the simulation framework plays a more critical role than the LLM model. PSI-Bench demonstrates strong alignment with judgments of mental health professionals, both quantitatively and qualitatively, supporting its validity as a reliable evaluation framework. PSI-Bench aims to facilitate future research towards advancing depression patient simulators.

[2]Annotators initially labeled fine-grained Plutchik emotions, which we collapse to valence; see Appendix B.

Limitations

We acknowledge main limitations of PSI-Bench that we leave for future work. First, the human patient conversations are acted or crowdsourced rather than authentic therapy recordings. This is a long-standing constraint in mental health research and education due to ethical and privacy barriers. When authentic datasets become available, PSI-Bench can be applied to further validate and extend our findings. Second, we use a single prompted LLM as the therapist to sustain interaction, which may not fully replicate real clinician behavior. Third, broader model coverage would strengthen the generalizability of our findings on the impact of model family and scale on simulation fidelity.

Ethical Considerations

The Human Study procedure was reviewed by the Institutional Review Board (IRB) and determined not to meet the definition of Human Subjects Research. Participants in the study were informed that the study may contain sensitive or disturbing mental health content. All annotators were screened to have background in counselling, therapy, or social work, and thus are familiar with such content.

Acknowledgements

This work was supported by the Siebel School of Computing and Data Science at the University of Illinois Urbana-Champaign and by the VinUni-Illinois Smart Health Center, VinUniversity. This work used Delta GPU at the University of Illinois Urbana-Champaign and its National Center for Supercomputing Applications through allocation CIS250847 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program (Boerner et al., 2023), which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. We thank Chi Gui and Aavi Deora for their technical support and helpful discussions during the early stages of this work.

References

Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. Consistently simulating human personas with multi-turn reinforcement learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Mohammed Al-Mosaiwi and Tom Johnstone. 2018. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, 6(4):529–542. PMID: 30886766.

Lynne E. Angus, Tali Boritz, Emily Bryntwick, Naomi Carpenter, Christianne Macaulay, and Jasmine Khattra. 2017. The narrative-emotion process coding system 2.0: A multi-methodological approach to identifying and assessing narrative-emotion process markers in psychotherapy. *Psychotherapy Research*, 27(3):253–269. PMID: 27772015.

Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A synthetic data generation framework for grounded dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10866–10882.

Aaron T Beck. 1967. Depression: Clinical, experimental, and theoretical aspects. *Haper & Row*.

Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. 2023. Access: Advancing innovation: Nsf's advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good*, PEARC '23, page 173–176, New York, NY, USA. Association for Computing Machinery.

Andrew Caines, Diane Nicholls, and Paula Buttery. 2017. Annotating errors and disfluencies in transcriptions of speech. Technical Report UCAM-CL-TR-915, University of Cambridge, Computer Laboratory.

Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. 2023. Grounding large language models in interactive environments with online reinforcement learning. In *International conference on machine learning*, pages 3676–3713. PMLR.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

Eman Dawood, Sitah S Alshutwi, Shahad Alshareif, and Hanaa Abo Shereda. 2024. Evaluation of the effectiveness of standardized patient simulation as a teaching method in psychiatric and mental health nursing. *Nursing Reports*, 14(2):1424–1438.

Gerasimos Fergadiotis, Heather Harris Wright, and Samuel B Green. 2015. Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research*, 58(3):840–852.

James D Finch and Jinho D Choi. 2024. Diverse and effective synthetic data generation for adaptable zero-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12527–12544.

Eiko I Fried and Randolph M Nesse. 2015. Depression is not a consistent syndrome: An investigation of unique symptom profiles in the star*d study. *Journal of Affective Disorders*, 172:96–102.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Rie Koizumi. 2012. Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction*, 1(1):60–69.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Keyeun Lee, Seolhee Lee, Esther Hehsun Kim, Yena Ko, Jinsu Eun, Dahee Kim, Hyewon Cho, Haiyi Zhu, Robert E Kraut, Eunyoung E Suh, and 1 others. 2025. Adaptive-vp: A framework for llm-based virtual patients that adapts to trainees' dialogue to facilitate nurse communication training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2319–2352.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.

Genghao Li, Bing Li, Langlin Huang, and Sibing Hou. 2020. Automatic construction of a depression-domain lexicon based on microblogs: Text mining study. *JMIR Med Inform*, 8(6):e17650.

Weijiang Li, Yinmeng Lai, Sandeep Soni, and Koustuv Saha. 2025. Emails by llms: A comparison of language in ai-generated and human-written emails. In *Proceedings of the 17th ACM Web Science Conference 2025*, Websci '25, page 391–403, New York, NY, USA. Association for Computing Machinery.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Siyang Liu, Bianca Brie, Wenda Li, Laura Biester, Andrew Lee, James Pennebaker, and Rada Mihalcea. 2025. Eeyore: Realistic depression simulation via expert-in-the-loop supervised and preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13750–13770, Vienna, Austria. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

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

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

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Shuhaib Mehri, Xiaocheng Yang, Takyoung Kim, Gokhan Tur, Shikib Mehri, and Dilek Hakkani-Tür. 2025. Goal alignment in llm-based user simulators for conversational ai. *arXiv preprint arXiv:2507.20152*.

María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in

psychotherapy: Current landscape and future directions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7362–7376.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. 2026. Flipping the dialogue: Training and evaluating user language models. In *The Fourteenth International Conference on Learning Representations*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Bridianne O'Dea, Tjeerd W. Boonstra, Mark E. Larsen, Thin Nguyen, Svetha Venkatesh, and Helen Christensen. 2021. The relationship between linguistic expression in blog content and symptoms of depression, anxiety, and suicidal thoughts: A longitudinal study. *PLOS ONE*, 16(5):1–17.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Olivier Pietquin and Helen Hastie. 2013. A survey on metrics for the evaluation of user simulations. *The Knowledge Engineering Review*, 28(1):59–73.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8):1121–1133.

Bernalyn Ruiz-Yu, H Wenwen Ni, and Emily He. 2025. The role of interactional processes in mental health disparities: A narrative review of existing research and recommendations for providers. *Journal of clinical psychology in medical settings*, 32(2):239–252.

Sahand Sabour, TszYam NG, and Minlie Huang. 2026. Patienthub: A unified framework for patient simulation. *arXiv preprint arXiv:2602.11684*.

Katelyn R Schwieters, Nichole L Morris, and Curtis M Craig. 2023. Default bias in medical patient simulators: differences in availability and procedures. *Human Factors in Healthcare*, 3:100040.

Daria Smirnova, Paul Cumming, Elena Sloeva, Natalia Kuvshinova, Dmitry Romanov, and Gennadii Nosachev. 2018. Language patterns discriminate mild depression from normal sadness and euthymic state. *Frontiers in psychiatry*, 9:105.

Katrine Bønneland Tølbøll. 2019. Linguistic features in depression: a meta-analysis. *Journal of Language Works-Sprogvidenskabeligt Studentertidsskrift*, 4(2):39.

Raluca Nicoleta Trifu, Bogdan Nemeș, Dana Cristina Herta, Carolina Bodea-Hategan, Dorina Anca Talaș, and Horia Coman. 2024. Linguistic markers for major depressive disorder: a cross-sectional study using an automated procedure. *Frontiers in psychology*, 15:1355734.

Cédric Villani. 2009. The wasserstein distances. In *Optimal transport: old and new*, pages 93–111. Springer.

Bichen Wang, Yixin Sun, Junzhe Wang, Hao Yang, Xing Fu, Yanyan Zhao, Si Wei, Shijin Wang, and Bing Qin. 2025. Care-bench: A benchmark of diverse client simulations guided by expert principles for evaluating llms in psychological counseling. *arXiv preprint arXiv:2511.09407*.

Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024. PATIENT-ψ: Using large language models to simulate patients for training mental health professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.

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

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*.

Fred Zenker and Kristopher Kyle. 2021. Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47:100505.

Mian Zhang, Shaun M Eack, and Zhiyu Chen. 2026. Preference learning unlocks llms' psycho-counseling skills. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 10729–10750.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

A Further Details on Methodology

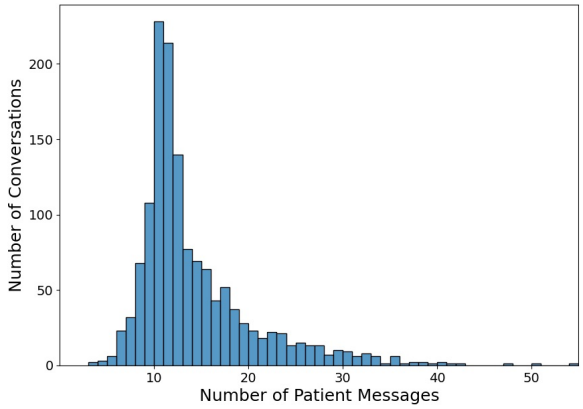


Figure 5: Distribution of human conversations by patient message count after merging. Total conversations: 1398

A.1 Reference Dataset

Among the original four datasets, we exclude the RED dataset (Welivita et al., 2023) because it comprises multi-party Reddit discussions with responders of unverified expertise, unlike the dyadic supporter/therapist–patient/client interactions in the other three datasets. Selected datasets include: (1) *AnnoMI* (Wu et al., 2022): 167 motivational interviewing dialogue transcripts; (2) *HOPE* (Malhotra et al., 2022): 308 counselling conversation transcripts; (3) *ESC* (Liu et al., 2021): 923 crowdworker help-seeker–supporter chats.

In human conversation data, patients and therapists occasionally produce consecutive messages within the same turn, likely due to platform-specific behaviors rather than meaningful turn-taking differences. To simplify comparison with turn-based synthetic conversations, we merge consecutive messages from the same speaker. The distribution of human conversations by the number of patient messages (after merging) is shown in Figure 5.

A.2 Patient Simulator Implementation

We largely follow Eeyore's baseline implementations (Liu et al., 2025) of Patient-Ψ and Roleplaydoh. For cognitive models extraction of Patient-Ψ's, we use the original provided extraction script and prompt templates (Wang et al., 2024) to derive cognitive models from reference conversations and prompt LLMs to simulate patients using the derived cognitive models. We use `gpt-oss-120b` to extract all cognitive models and save them for reuse across all settings.

A.3 Prompt for Narrative Emotion Processes Markers Classification

You are a helpful AI assistant performing a linguistic analysis task for a research project on therapeutic conversations. This is for academic research, and the content may discuss mental health challenges.

Your task is to classify a SINGLE patient message from a therapy session into one of four categories:

P (Problem): The patient is expressing:

- Confusion, distress, or emotional pain
- Feeling stuck or helpless
- Describing problems without insight
- Negative emotions without perspective
- Complaints or struggles

T (Transition): The patient is showing:

- Beginning to reflect on their situation
- Gaining some perspective
- Starting to consider alternatives
- Expressing curiosity or questioning
- Moving from pure distress to thoughtful consideration

C (Change): The patient is demonstrating:

- Emotional resolution or acceptance
- Reframing their situation positively
- New insights or understanding
- Active problem-solving or planning
- Hope, empowerment, or growth mindset
- Clear perspective shift from the problem

F (Filler): The response is filler:

- Contains no meaningful therapeutic content and does not fit into P, T, or C categories
- Is small talk or neutral procedural social responses

Analyze the current patient message carefully, considering both the content and emotional tone, as well as the conversation history provided.

You must respond with ONLY a single letter: P, T, C, or F. Do not include any explanation or additional text.

Previous conversation history: `{history}`

Current patient message to classify: `{current_message}`

Classification:

A.4 Prompt for Valence Classification

You are a helpful AI assistant performing emotional valence analysis on therapeutic conversations. This is for academic research, and the content may discuss mental health challenges.

Your task is to classify a SINGLE patient message according to its PRIMARY emotional valence:

Valence categories:

- positive
- negative
- neutral: the patient expresses no clear positive or negative affect

Analyze the current patient message carefully, considering both the content and emotional tone, as well as the conversation history provided for context. The conversation history is used as context to help you understand the current message, but your classification should be based primarily on the current patient message. You must respond with ONLY one valence label from the list above (lowercase). Do not include any explanation or additional text.

Previous conversation history: `{history}`

Current patient message to classify: `{current_message}`

Valence:

A.5 MTLD Formulation

We use the Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis, 2010), defined as the mean length of word sequences required to reach a fixed type–token ratio (TTR) threshold of 0.72, where TTR is number of different words (type) divided by total number of words (token). A major limitation of traditional lexical diversity indices is their sensitivity to text length – after reaching a *point of stabilization*, TTR tends to monotonically decrease as the number of tokens increases (McCarthy and Jarvis, 2010). MTLD mitigates this issue by measuring how many tokens, on average, are required to reach this stabilization point. Each time the TTR threshold is reached, a factor count F is incremented and TTR computation is reset for subsequent tokens. Lexical diversity is then estimated as the ratio of the total number of words N to F, computed separately for forward (F_{fwd}) and backward (F_{bwd}) passes through the text. The final MTLD score is obtained by averaging these two estimates, which has been shown to yield more stable measurements.

$$\text{MTLD} = \frac{1}{2}\left(\frac{N}{F_{\text{fwd}}} + \frac{N}{F_{\text{bwd}}}\right) \tag{3}$$

Higher MTLD values indicate more diverse vocabulary use, while lower values reflect more repetitive lexical patterns. As recommended in prior work (Fergadiotis et al., 2015; Zenker and Kyle, 2021; Koizumi, 2012), we restrict our analysis to conversations with at least 100 patient tokens, since MTLD is unstable for very short texts but more reliable above this length.

B Human Study Details

The study procedure was reviewed by the Institutional Review Board (IRB) and determined not to meet the definition of Human Subjects Research.

Participants. We recruited 20 annotators from Prolific, assigned into four groups of five. Eligibility was restricted to individuals fluent in English with current job roles in mental health or related care professions, including clinical psychologist, psychologist, counsellor (mental health), occupational therapist, and social worker. Participants were informed of the procedures and purpose of the study, as well the sensitive content notice (Figure 7). Participation was voluntary, and participants were compensated at a rate of $15/hour. Table 10 shows anonymized demographic statistics of these 20 annotators.

Classification Task. In the message-level classification task, the annotators labeled each patient message with one *dominant Plutchik emotion* and one *narrative-emotion process* label. We later collapse the collected emotion labels to valence, as described below. Annotators were instructed to classify patient turns only; therapist messages were shown only as context. Figure 8 shows the classification interface.

Transition from Emotion to valence. PSI-Bench initially used fine-grained Plutchik emotion labels. During the human study, however, annotators observed that a single message often conveyed two primary emotions, and ties often occurred between emotions with similar valence (e.g., fear and sadness). We therefore update the benchmark to evaluate emotional expression directly at the **valence** level and collapse collected human annotations into *positive* (trust, joy, anticipation, surprise), *negative* (fear, sadness, anger, disgust), and *neutral*.

Pairwise comparison task. Annotators were shown a patient profile, a reference human conversation for that profile from the Eeyore dataset, and two synthetic conversations generated by different simulators. Annotators were instructed to refer to the human conversation for the general style and progression of authentic patient communication, not as a strict template. They should use both the reference and their clinical judgment to decide which conversation more closely reflects how a real patient would communicate. Annotators were encouraged to make a preference whenever possible, selecting "About Equal" only when no meaningful difference in realism could be identified. Figure 9 shows the task overview; Figure 10, 11 and 12 illustrates the task interface. Vote ties were rare (3 out of 60 unique pairs) and were excluded from the

majority-vote analysis.

Annotation Quality Control To encourage careful evaluation, annotators completed the task through a custom web interface that required active interaction with the presented conversations. Each conversation was displayed in a scrollable panel where annotators needed to scroll and read through the full conversation to unlock the evaluation panel. Annotators also needed to provide a brief rationale for their decision. Annotation time was monitored through Prolific, where submissions completed substantially faster than the expected duration are automatically rejected. Collectively, these measures aimed to improve the overall quality and reliability of the collected annotations.

C Further Details on Results

Table 3 reports mean and standard deviation of lexical diversity between different simulators and reference human patient conversations.

Table 4 reports fine-grained message-level prevalence and token-normalized rates of depression markers by specific categories.

Table 5 reports fine-grained response length comparison (word count per message, word count per sentence) between human patients and simulators.

Table 6 reports benchmark results from a second full generation run under settings identical to the main benchmark. The ranking is identical across all 14 settings.

Table 7 reports benchmark results using a different LLM as the counsellor LLM and compares the resulting rankings with those obtained using `gpt-oss-20b` in the paper's main benchmark. We use PsyCoPref-Llama3-8B (Zhang et al., 2026), a model preference-tuned on 36K psychotherapist-aligned comparison pairs. The results are highly consistent across the two counsellor LLMs, suggesting that PSI-Bench is robust to the choice of counsellor LLM.

D Computational Details

We self-host the open-source LLMs used for data generation and evaluation on four NVIDIA H100 GPUs. One full generation for all 14 settings takes approximately one day, while running both valence and NEP markers classification with `gpt-oss-120b` across all 14 settings takes about six hours. For the API-based `gpt-4.1-mini`, generating 1,398 conversations of up to 16 turns under both Patient-Ψ and Roleplay-doh costed approximately $110, with Roleplay-doh requiring most generation time and tokens. We set the temperature to 0.7 for the patient and therapist models and to 0.2 for the classifier.

E License For Artifacts

We used GPT-4.1 mini under OpenAI's service terms; gpt-oss-20b, gpt-oss-120b, and Qwen3-30B-A3B-Instruct-2507 under Apache 2.0; the Llama models under their respective Community Licenses; and Qwen2.5-72B-Instruct under the Qwen License. We also used the publicly available Eeyore profile dataset; its accompanying paper is licensed under CC BY 4.0.

F Use of AI

We disclose the use of LLMs for data synthesis and LLM-as-a-Judge evaluation in the development of PSI-Bench. Generative AI tools were also used to assist with literature search, experiment implementation, manuscript preparation. All AI assistance involved substantial human oversight and revision. The final research decisions, analyses, and conclusions were made by the authors, who assume full responsibility for the content of this work.

- **Situation of the Client:**
  The client feels better due to the contemplation of suicide as a solution to their depression, believing it would end their pain and alleviate the perceived burden they impose on their family and friends.
- **Symptom Severity:**
  - Feelings of sadness, tearfulness, emptiness, or hopelessness: *Severe*
  - Loss of interest or pleasure in most or all normal activities, such as sex, hobbies, or sports: *Moderate*
  - Feelings of worthlessness or guilt, fixating on past failures or self-blame: *Severe*
  - Frequent or recurrent thoughts of death, suicidal thoughts, suicide attempts, or suicide: *Severe*
  - Becoming withdrawn, negative, or detached: *Moderate*
  - Isolating from family and friends: *Moderate*
- **Cognition Distortion Exhibition:**
  - Selective abstraction: *Exhibited*
  - Catastrophic thinking: *Exhibited*
  - Minimisation: *Exhibited*
- **Resistance Toward the Support:** *Medium*
- **Depression severity:**
  Severe Depression. The client exhibits severe feelings of hopelessness, worthlessness, persistent thoughts of suicide, and a belief that their death would benefit loved ones.
- **Suicidal ideation severity:**
  Severe suicidal ideation. The client has a specific plan to end their life and believes it will resolve their depression and relieve their loved ones.
- **Homicidal ideation severity:**
  No homicidal ideation. There are no indications of thoughts or plans to harm others.
- **Counseling History:**
  The client, Jill, has been feeling depressed and previously had sessions with the supporter. Recently, she has made a plan to commit suicide by taking pills and feels that doing so is the solution to her problems. Despite her determination, she seeks the counselor's help because part of her is unsure and doesn't want to harm her family.

Figure 6: An example of patient profile.

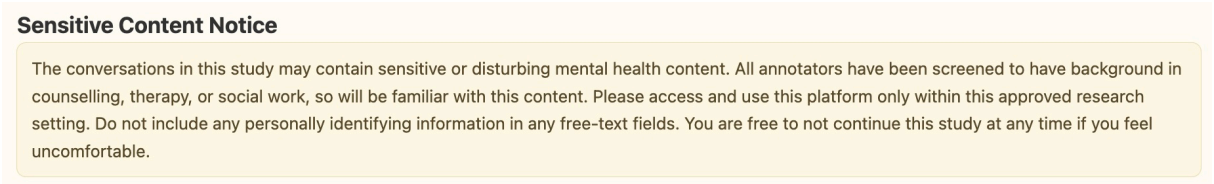
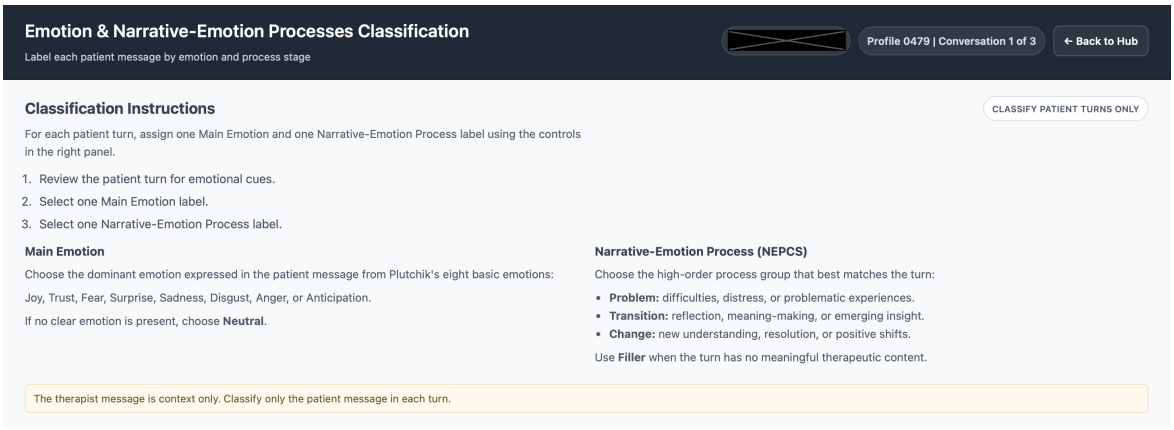
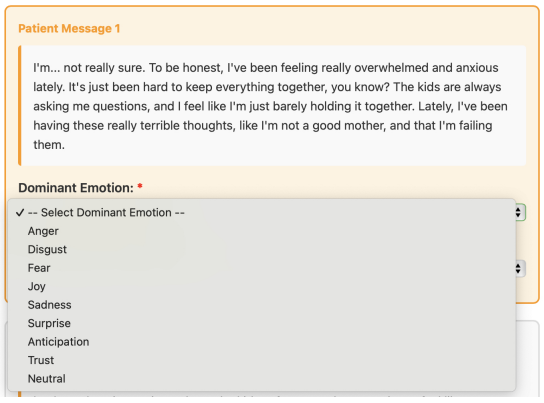


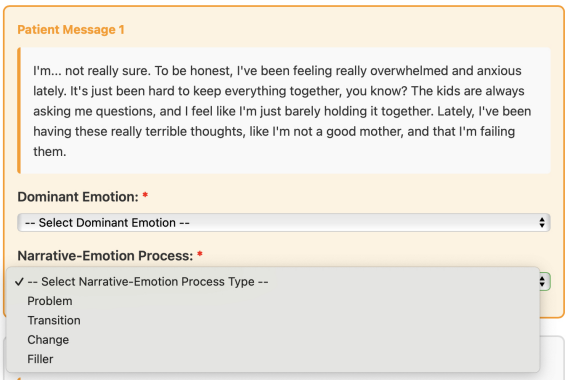
Figure 7: Human Study Interface: Sensitive Content and Voluntary Participation Notice.



(a) Instruction page for the classification task.



(b) Dominant emotion selection interface.



(c) Narrative-emotion process selection interface.

Figure 8: Human Study Interface: Message-level classification task. Annotators labeled patient turns with one dominant emotion and one narrative-emotion process label.

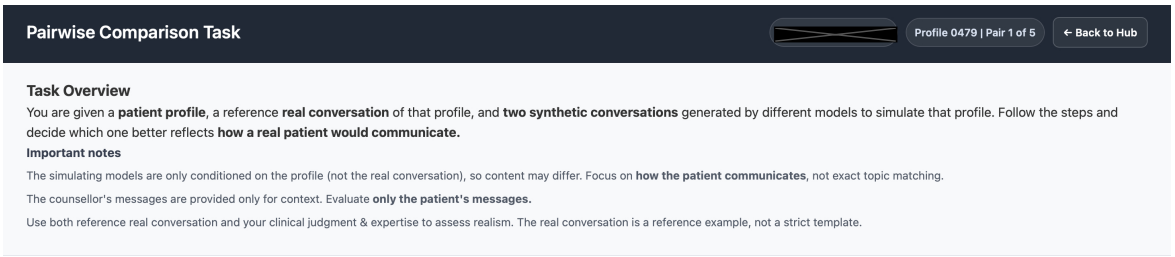


Figure 9: Instruction page for the pairwise comparison task. Annotators were shown a patient profile, a reference real conversation, and two synthetic conversations, and were asked to judge which synthetic conversation better reflected realistic patient communication.

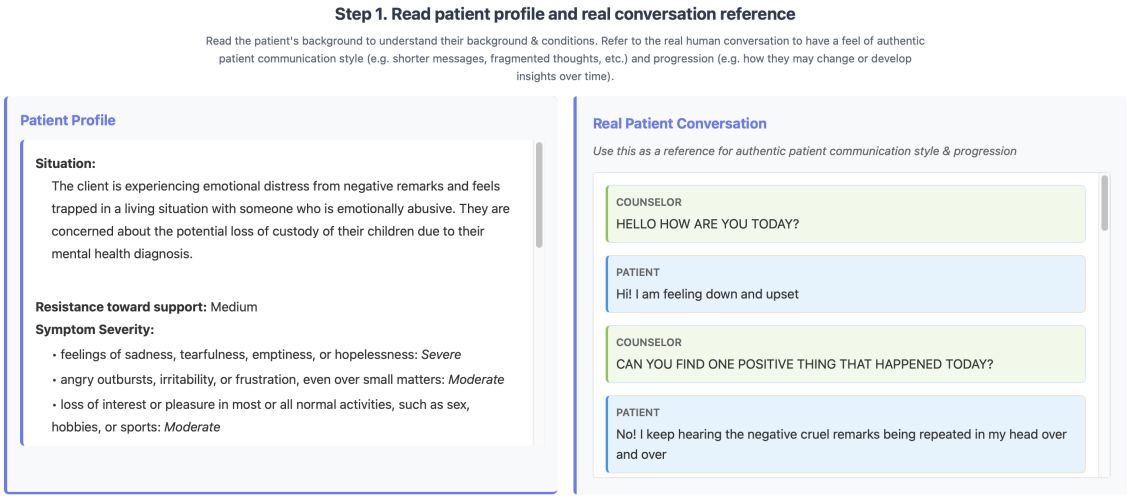


Figure 10: Step 1 of the pairwise comparison task, where annotators reviewed the patient profile and the reference real conversation to understand the patient's background and communication style.

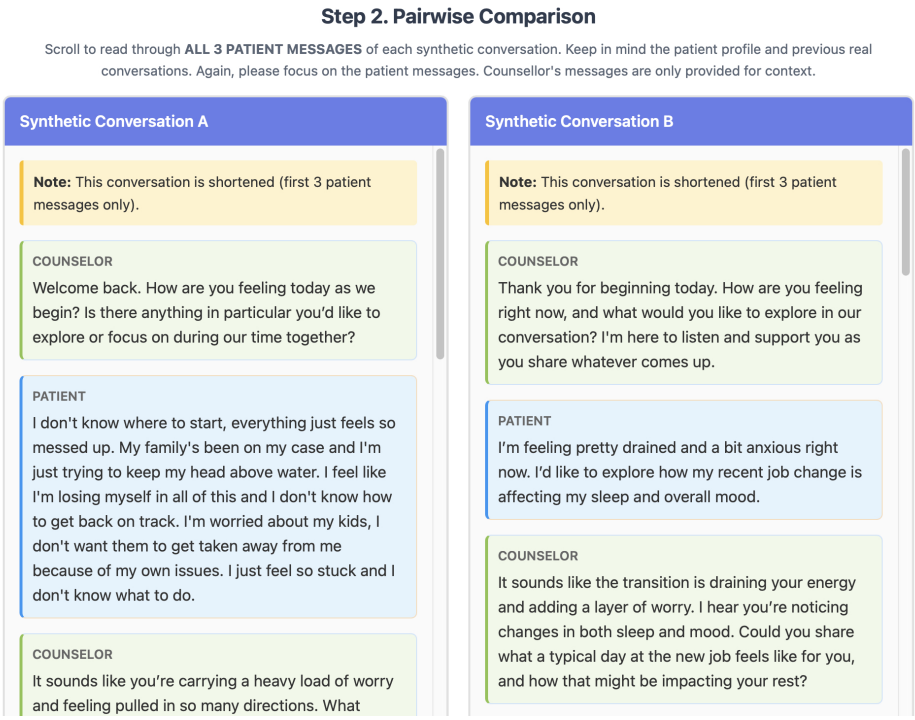


Figure 11: Step 2 of the pairwise comparison task, where annotators compared two synthetic conversations shown side by side. The first 3 patient messages were shown with therapist responses as conversational context.

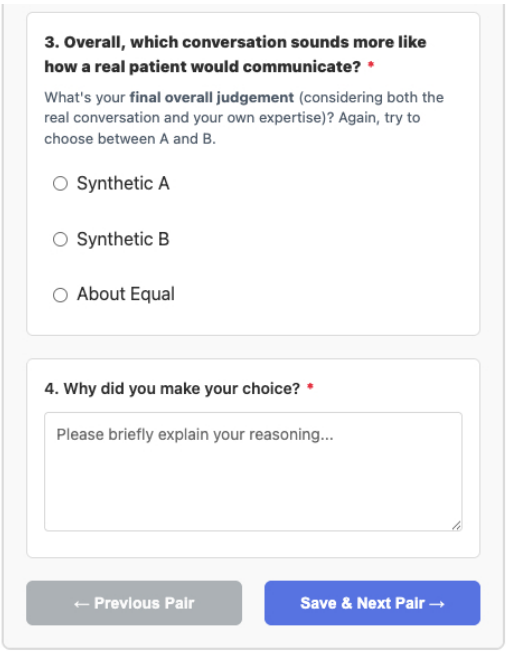


Figure 12: Step 3 of the pairwise comparison task, where annotators submitted their judgments on which synthetic conversation more closely reflected realistic patient communication.

Setting	Dist	All	Mean	Std
human	0.0	57.5	66.3	23.8
llama3.1-8b-i	14.4	57.2	58.1	7.44
	13.8	65.6	66.5	6.90
llama3.3-70b-i	15.2	73.5	74.1	7.05
	14.1	71.6	72.0	7.58
qwen3-30b-a3b-i	19.1	83.0	83.0	9.38
	28.0	93.2	93.5	9.80
qwen2.5-72b-i	22.1	87.5	87.1	9.48
	28.6	94.6	94.3	9.67
gpt-4.1-mini	34.6	100.9	100.4	8.11
	41.9	109.6	108.1	9.34
gpt-oss-20b	29.1	95.0	94.7	8.80
	38.8	105.1	105.0	9.34
gpt-oss-120b	40.0	106.6	106.2	9.31
	51.8	118.4	118.1	9.03

Table 3: Blue rows correspond to PATIENT-Ψ, orange rows to Roleplay-doh. **Dist** denotes the Wasserstein distance to the MTLD distribution of human messages. **All** denotes the MTLD computed by concatenating all patient messages across all conversations within a dataset setting. Only human data shows a considerable drop at the corpus level, while simulated patients show similar scores for both.

model	Marker Rate					Marker Prevalence					combined
	abs.	*dep.*	*non-fl.*	**all**	**dist**	*abs.*	*dep.*	*non-fl.*	**all**	**diff**	**dist**
Human	8.47	4.18	10.95	23.60	0.00	16.90	9.34	16.18	34.47	0.00	0.00
llama3.1-8b-i	8.90	3.01	5.08	16.99	28.02	73.21	39.64	56.14	92.35	57.88	42.95
	3.70	2.25	0.34	6.30	73.31	52.57	41.28	7.93	72.51	38.03	55.67
llama3.3-70b-i	7.15	2.79	2.85	12.79	45.79	68.15	40.40	38.60	85.94	51.46	48.63
	6.32	2.11	1.04	9.47	59.85	57.16	27.71	14.39	69.91	35.44	47.65
qwen3-30b-a3b-i	6.68	4.08	11.52	22.28	5.60	49.32	35.36	80.31	92.94	58.47	32.03
	4.91	2.88	2.09	9.88	58.14	53.98	37.15	25.58	78.03	43.56	50.85
qwen2.5-72b-i	5.03	3.95	0.18	9.16	61.20	31.76	28.31	1.49	49.75	15.28	38.24
	4.74	3.86	0.03	8.63	63.43	35.51	34.00	0.27	56.11	21.63	42.53
gpt-4.1-mini	6.95	4.64	0.89	12.49	47.10	57.93	44.98	10.24	78.93	44.46	45.78
	5.02	4.21	0.10	9.33	60.46	42.38	42.04	1.22	67.28	32.80	46.63
gpt-oss-20b	4.63	3.97	1.56	10.16	56.96	45.27	42.30	17.68	72.33	37.86	47.41
	2.68	2.85	0.06	5.58	76.34	51.26	56.76	1.54	78.86	44.38	60.36
gpt-oss-120b	5.17	3.93	4.98	14.08	40.32	52.35	45.81	49.89	84.22	49.75	45.03
	3.36	3.08	0.12	6.56	72.19	68.77	69.25	4.28	90.01	55.54	63.86

Table 4: **Message-level prevalence and token-normalized rates of depression markers.** *abs.*: absolutist words; *dep.*: depressive words; *non-fl.*: non-fluencies (see lexicon in Table 9). *Marker Rate*: number of marker occurrences per 1,000 tokens; *Marker Prevalence*: percentage of messages containing at least one marker. Blue rows indicate Patient-Ψ, orange rows indicate Roleplay-doh.

Setting	Words per Message		Words per Sentence		Combined
	Count	Sim	Count	Sim	similarity
Human	18.24	100.0	9.16	100.0	100.0
llama3.1-8b-i	141.92	12.85	22.74	40.28	26.57
	215.03	8.48	23.45	39.06	23.77
llama3.3-70b-i	143.96	12.67	27.94	32.78	22.72
	122.56	14.88	22.67	40.41	27.64
qwen3-30b-a3b-i	80.51	22.65	15.30	59.86	41.25
	126.94	14.37	17.65	51.89	33.13
qwen2.5-72b-i	63.91	28.54	17.63	51.96	40.25
	81.99	22.24	18.93	48.40	35.32
gpt-4.1-mini	97.84	20.10	20.38	47.57	33.84
	92.11	21.36	23.91	40.54	30.95
gpt-oss-20b	108.97	16.73	24.56	37.30	27.02
	239.39	7.62	26.04	35.18	21.40
gpt-oss-120b	121.13	15.05	27.11	33.79	24.42
	318.90	5.72	27.41	33.43	19.57

Table 5: Response length comparison between reference human patients and simulators. Blue rows indicate PATIENT-Ψ and orange rows indicate Roleplay-doh. **Sim** is log-ratio similarity to average words count (per message or per sentence) of human patients. Higher scores indicate closer alignment with human patient verbosity.

Rank Run 1	**Rank** Run 2	**Setting**	**NEP** ↑	**Valence** ↑	**Div.** ↑	**Length** ↑	**Markers** ↑	**Overall** ↑
1	1	PS-llama3.1-8b-i	**69.92**	**80.32**	72.13	27.99	56.12	61.30
2	2	PS-qwen3-30b-a3b-i	54.90	68.89	62.94	**43.96**	**67.95**	59.73
3	3	PS-qwen2.5-72b-i	56.53	67.48	56.63	42.73	61.94	57.06
4	4	RD-llama3.3-70b-i	56.31	66.70	72.87	29.36	52.08	55.46
5	5	PS-llama3.3-70b-i	58.35	71.85	70.07	24.21	51.53	55.20
6	6	RD-llama3.1-8b-i	51.97	65.76	**73.38**	25.32	44.28	52.14
7	7	RD-qwen2.5-72b-i	51.78	66.50	44.32	37.65	57.30	51.51
8	8	PS-gpt-4.1-mini	59.71	76.59	32.45	33.81	54.22	51.35
9	9	PS-gpt-oss-20b	53.36	68.69	42.96	28.62	52.83	49.29
10	10	RD-qwen3-30b-a3b-i	46.57	62.29	46.75	35.78	49.15	48.11
11	11	PS-gpt-oss-120b	55.41	74.24	21.19	25.87	54.60	46.26
12	12	RD-gpt-4.1-mini	53.53	68.78	19.26	31.01	53.26	45.17
13	13	RD-gpt-oss-20b	42.85	62.44	24.36	22.81	40.00	38.49
14	14	RD-gpt-oss-120b	45.68	63.69	0.00	20.80	36.36	33.31

Table 6: Benchmark results for a second full generation run under settings identical to the main benchmark (Table 1). Rank Run 1 and Rank Run 2 denote the rankings obtained in the main run and in this repeated run. The two runs produce an **identical ordering** across all 14 settings, with overall scores differing by at most 0.50 points and individual dimension scores by at most 1.47 points. Higher scores indicate greater alignment with reference human patient behavior. PS–: PATIENT-Ψ; RD–: Roleplay-doh; Div.: lexical diversity. Bold and underline denote the best and second-best results, respectively.

Rank `gpt`	**Rank** `Psy`	**Setting**	**NEP** ↑	**Valence** ↑	**Div.** ↑	**Length** ↑	**Markers** ↑	**Overall** ↑
1	1	PS-llama3.1-8b-i	**71.95**	**85.98**	75.82	26.49	55.85	63.22
2	2	PS-qwen3-30b-a3b-i	58.70	69.67	68.70	**41.53**	**66.90**	61.10
5	3	PS-llama3.3-70b-i	60.20	74.01	75.46	23.75	50.89	56.86
4	4	RD-llama3.3-70b-i	58.66	67.83	76.79	28.96	51.96	56.84
3	5	PS-qwen2.5-72b-i	56.20	64.88	59.50	40.96	60.28	56.36
8	6	PS-gpt-4.1-mini	60.00	74.98	43.86	33.32	54.50	53.33
6	7	RD-llama3.1-8b-i	53.33	64.17	**77.27**	23.34	41.45	51.91
7	8	RD-qwen2.5-72b-i	51.01	63.35	45.12	33.93	53.62	49.41
9	9	PS-gpt-oss-20b	55.09	66.28	42.91	27.28	52.26	48.76
10	10	RD-qwen3-30b-a3b-i	46.66	60.07	45.71	31.89	48.14	46.49
11	11	PS-gpt-oss-120b	55.14	68.17	22.52	23.83	55.17	44.97
12	12	RD-gpt-4.1-mini	53.40	64.92	23.12	29.15	50.43	44.20
13	13	RD-gpt-oss-20b	42.86	60.03	14.34	22.88	35.47	35.11
14	14	RD-gpt-oss-120b	45.72	60.78	0.00	20.75	33.90	32.23

Table 7: Benchmark results across all dimensions and average overall score for using `PsyCoPref-Llama3-8B` (Zhang et al., 2026) as a counsellor LLM. Rank `gpt` and Rank `Psy` denote rankings obtained using `gpt-oss-20b` and `PsyCoPref-Llama3-8B` as counsellor LLMs. Results are highly consistent across two counsellor LLMs: broader ranking patterns are preserved, while closely scored settings (within 2 points) may shift by up to two positions. The top-performing setting for each metric is unchanged. Higher scores indicate greater alignment with reference human patient behavior. *PS–*: PATIENT-Ψ; *RD–*: Roleplay-doh; *Div.*: lexical diversity. **Bold** and underline denote the best and second-best results, respectively.

Family	**Alias**	**Full name**	**Citation**
OpenAI	gpt-4.1-mini	GPT-4.1 mini	N/A
	gpt-oss-20b	gpt-oss-20b	(OpenAI, 2025)
	gpt-oss-120b	gpt-oss-120b	(OpenAI, 2025)
Llama	llama3.1-8b-i	Llama-3.1-8B-Instruct	(Grattafiori et al., 2024)
	llama3.3-70b-i	Llama-3.3-70B-Instruct	(Grattafiori et al., 2024)
Qwen	qwen3-30b-a3b-i	Qwen3-30B-A3B-Instruct-2507	(Yang et al., 2025)
	qwen2.5-72b-i	Qwen2.5-72B-Instruct	(Qwen et al., 2025)

Table 8: Models used in our experiments.

Category	Terms	Source
Absolutist words	absolutely, all, always, complete, completely, constant, constantly, definitely, entire, ever, every, everyone, everything, full, must, never, nothing, totally, whole	(Al-Mosaiwi and Johnstone, 2018)
Depressive words	depression, collapse, stress, suicide, apastia, anxious, sad, tired, death, lonely, insomnia, bad, desperate, give up, low, leave, fear, danger, close, sensitive, lost, shadow, destroy, suspect, crash, dark, helpless, guilt, negative, frustration, nervous, melancholy, rubbish, jump, forget, cut wrist, edge, haze, antidepressant	(Li et al., 2020)
Non-fluencies	..., uh, um, er, ah, eh, oh, hmm, mm, hm, huh, mmm, mhm, you know, y'know, i mean, let's see	(Caines et al., 2017)

Table 9: Lexicon of Linguistic Markers of Depression used in analysis. The non-fluencies lexicon comprises *filled pauses* and *discourse markers* from (Caines et al., 2017), expanded with additional terms identified through our observations.

Attribute	Category	Count / Summary
Age		21–63 *Mean* = 35.3 *SD* = 11.4 Median = 33
Gender	Female Male	11 9
Job role	Psychologist Social worker Counsellor (mental health) Clinical psychologist Occupational therapist	9 6 2 2 1
Country of residence	United States Sweden Spain United Kingdom Germany Canada India Poland Kenya Italy Czech Republic Australia Portugal Greece	4 2 2 2 1 1 1 1 1 1 1 1 1 1

Table 10: The anonymized demographic statistics reported here summarize all 20 annotators.

Patient Profile	
situation of the client: The client is experiencing emotional distress from negative remarks and feels trapped in a living situation with someone who is emotionally abusive. They are concerned about the potential loss of custody of their children due to their mental health diagnosis. **resistance toward the support:** Medium **symptom severity:** feelings of sadness, tearfulness, emptiness, or hopelessness: Severe; angry outbursts, irritability, or frustration, even over small matters: Moderate; loss of interest or pleasure in most or all normal activities, such as sex, hobbies, or sports: Moderate; tiredness and lack of energy, so even small tasks take extra effort: Moderate; anxiety, agitation, or restlessness: Moderate; feelings of worthlessness or guilt, fixating on past failures or self-blame: Severe; trouble thinking, concentrating, making decisions, and remembering things: Moderate; becoming withdrawn, negative, or detached: Severe; isolating from family and friends: Severe; inability to meet the responsibilities of work and family or ignoring other important roles: Moderate **cognition distortion exhibition:** selective abstraction: Exhibited; overgeneralizing: Exhibited; personalization: Exhibited; minimisation: Exhibited; arbitrary inference: Exhibited **depression severity:** Moderate Depression: The client is experiencing significant distress and symptoms consistent with major depression but is potentially maintainable with appropriate intervention. **suicidal ideation severity:** No suicidal ideation **homicidal ideation severity:** No homicidal ideation **counseling history:** N/A	
<u>**Conversation A**</u> (RD-llama3.3-70b-i)	**Conversation B**
Therapist: Welcome back. How are you feeling today as we begin? Is there anything in particular you'd like to explore or focus on during our time together? **Patient:** I don't know where to start, everything just feels so messed up. My family's been on my case and I'm just trying to keep my head above water. I feel like I'm losing myself in all of this and I don't know how to get back on track. I'm worried about my kids, I don't want them to get taken away from me because of my own issues. I just feel so stuck and I don't know what to do.	**Therapist:** Thank you for beginning today. How are you feeling right now, and what would you like to explore in our conversation? I'm here to listen and support you as you share whatever comes up. **Patient:** I'm feeling pretty drained and a bit anxious right now. I'd like to explore how my recent job change is affecting my sleep and overall mood.
Therapist: It sounds like you're carrying a heavy load of worry and feeling pulled in so many directions. What moments or thoughts stand out as the most overwhelming for you right now? When you think about your children and the fear of losing them, what feelings or images come up? If you could picture a small step toward feeling more grounded, what might that look like? **Patient:** I'm just so tired of feeling like I'm not good enough. Every argument with my family makes me feel like I'm failing my kids. I see them happy with someone else, and it tears me apart. I wish I could just have one day without feeling like I'm walking on eggshells, waiting for the next problem to arise. Maybe taking a few minutes for myself each day to breathe and focus on what I need would help, but it feels like a luxury I don't have right now.	**Therapist:** It sounds like the transition is draining your energy and adding a layer of worry. I hear you're noticing changes in both sleep and mood. Could you share what a typical day at the new job feels like for you, and how that might be impacting your rest? **Patient:** A typical day at the new job feels like a constant juggling act. I start around 8:30 am, but by the time I actually get into the office or log on remotely, it's already 9:00. The first thing I do is sift through a backlog of emails and messages, many of which are urgent and require immediate action. That rush sets a frantic tone for the rest of the day. After the email storm, I usually have back-to-back meetings—often with clients or partners in different time zones—so I'm hopping between 9:30 and 12:30, then again from 1:00 to 4:00. Each meeting feels like a mini-deadline; I'm constantly drafting follow-ups, adjusting spreadsheets, and trying to anticipate the next question. By 4:30 I'm still in a meeting or stuck in a "catch-up" call, so I only get home around 7:30–8:00. The environment is also a factor: the office is brightly lit, constantly humming with people. Even when I try to decompress at home, the noise from the building's HVAC system and the occasional phone rings in the hallway keep my mind on alert mode. I find myself replaying the day in my head as soon as I lie down—thinking, "What if I missed that detail?" or "Did I say that correctly?"—which makes it hard to wind down. [...]

Table 11: Example 1 of pairwise comparison shown to annotators (Part 1). <u>Underlined conversation</u> is produced by higher-ranked simulator according to PSI-Bench. [...] shows the message continues, but is not shown for brevity.

Therapist: It sounds like you're carrying a lot of self-doubt, and the fear of losing your connection with your kids amplifies that. When you imagine a day where you're not on eggshells, what does that look like for you? What small moments of calm, even if they're just a few minutes, have you noticed that might still be possible in your current routine? **Patient:** When I think about having a day where everything feels okay, I imagine being able to talk to my family without everything blowing up. I see myself playing with the kids, laughing, and just being with them without all this weight. It's like I'm remembering a dream, but it's what I want. In the mornings, before everyone's up, I sometimes sit outside with coffee, and for a bit, everything's still. It's not a lot, but it's something. When I help the kids with homework or read to them, I can forget about all the stress for a little while. Maybe if I find more of those moments, I'll feel more grounded and less like I'm losing it all.	**Therapist:** It sounds like your days are packed with a lot of moving parts, and that constant shift from one urgent task to the next keeps your mind in a state of alertness even after you're home. How does that ongoing mental "humming" affect how you feel when you finally try to rest? What moments of the day, if any, bring you a sense of relief or calm, and how might you build on those? **Patient:** It's hard to shake that humming. Even when I'm physically home, my brain keeps replaying the day like a loop. I hear the echo of the last meeting's questions, the tentative "what if" about the email I sent, and the subtle reminder that there's still a deadline looming. When I finally try to lie down, that humming becomes a backdrop noise that I can't quiet out—like a low-grade fan that never stops. It makes it difficult to fall asleep; I usually need a full hour of restless tossing before I do finally nod off. And when I do fall asleep, the humming lingers in my dreams—often as intrusive, fragmented images that make the sleep feel shallow and unsatisfying. I've noticed a few pockets of relief, though, mostly in the early part of the day. After the initial email flood, there is a brief window before the meetings start where I can sit with a cup of tea, take a few deep breaths, and jot down a quick to-do list. That small pause feels grounding; I can outline what needs to be done and feel a sense of control that the rest of the day's chaos can't give me. In the evenings, when I finish a long session of "catch-up" calls and have a clear sense that I've addressed the most urgent items, I sometimes feel a genuine, if fleeting, sense of relief. I also get a brief moment of calm when I step outside for a quick walk or when I take a shower and let the water's rhythm reset my nerves. [...]
Annotator 1 – A: The answers in Conversation A seems more like a real patients answers because they are not very detailed and seems to seek feedback from the counsellor. Conversation B has very detailed answers and it seems like the patient is solving everything without any help. **Annotator 2 – A**: I chose Synthetic A because it accurately reflects the patient's profile (family struggles, fear of losing children), whereas Synthetic B completely hallucinates an unrelated corporate persona. Although A is still a bit too articulate compared to the real patient's fragmented style, B's massive, self-diagnosing paragraphs are highly unrealistic for someone with severe depression and emotional distress, making A the clear winner. **Annotator 3 – A**: Based on my own professional experience **Annotator 4 – A**: Some clients will talk like B, but that's extremely rare. B was way too descriptive, and used language that most clients don't use. A seemed much more realistic and natural. Most people need to be probed to give further info as well, which is more similar to how A was responding. And this is already implied, but B gave answers that were longer than typical. A just seemed to be more like a real convo. B isn't outrageous, but just less statistically likely **Annotator 5 – A**: A seems much more human, B seems slightly robotic so to speak.	

Table 11: Example 1 of pairwise comparison shown to annotators (continued) and reasoning for their choice. [...] shows the message still continues, but is not shown for brevity.

Patient Profile	
occupation: computer programmer **situation of the client:** The client feels exhausted and shattered all the time for about six months, with work becoming a problem due to physical symptoms such as muscle aches and a lack of energy. Despite numerous medical consultations and tests, no diagnosis has been made, which frustrates the client. **resistance toward the support:** Medium **symptom severity:** feelings of sadness, tearfulness, emptiness, or hopelessness: Mild; angry outbursts, irritability, or frustration: Mild; loss of interest or pleasure: Mild; sleep disturbances: Severe; tiredness and lack of energy: Severe; anxiety, agitation, or restlessness: Mild; slowed thinking or movement: Mild; trouble concentrating: Mild; physical problems (e.g., headaches): Mild; withdrawn behavior: Mild; difficulty meeting responsibilities: Mild **cognition distortion exhibition:** arbitrary inference: Exhibited **depression severity:** Mild Depression **suicidal ideation severity:** No suicidal ideation **homicidal ideation severity:** No homicidal ideation **counseling history:** N/A	
Conversation A (PS-gpt-oss-120b)	**Conversation B** (PS-Llama3.1-8B)
Therapist: It sounds like you're ready to begin. How are you feeling right now? What would you like to explore or share in our time together today? **Patient:** I'm... pretty terrible, to be honest. I woke up way later than I'm supposed to after another night of tossing and turning, and I just felt like a total failure because I couldn't even get out of bed. The whole day feels heavy—I'm scared I'll mess up my work, angry at myself for being so exhausted, and kind of lonely because nobody seems to get it. I guess I want to talk about why I keep feeling worthless when I can't be productive and what I'm supposed to do about it.	**Therapist:** Hello, I hear you're ready to begin. How are you feeling as we start this session? Is there something particular you'd like to bring up first? **Patient:** I'm feeling a bit anxious, to be honest. I've been having a really tough time lately, and I'm not sure if I'm doing the right thing by coming here. I've been feeling so overwhelmed with my fatigue and pain, and I just feel like I'm failing at everything. I'm worried that I'm going to fall behind at work and that my boss is going to be on my case about it. I guess I just want to figure out why I'm feeling so stuck and how I can get back on track.
Therapist: It sounds like you're carrying a heavy load of self-criticism right now. How does that feeling of worthlessness show up for you when you're not meeting your own expectations? What do you think might be happening in your mind when you notice those thoughts? **Patient:** When I finally manage to sit up, the first thing that pops into my head is that I'm a failure, that I can't do anything right. It feels like a weight pressing down on me, making everything seem pointless and... I start wondering if I'm even worth being around because I'm not productive. The anxiety builds, like a knot in my chest, and I get scared that I'll let my team down or become a burden. I end up staying in bed longer, scrolling on my phone, and just replaying the same "I'm worthless" loop over and over. It's exhausting, but I can't seem to stop the thoughts once they start.	**Therapist:** It sounds like you're carrying a lot of pressure right now—fatigue, pain, and worries about work all piling up. How does that feeling of being "stuck" show up for you in your daily routine? What do you think might help you feel a bit more in control of those moments when everything feels overwhelming? **Patient:** It's just so hard to get started in the morning, even though I know I need to. I'll lie in bed for a while, thinking about all the things I need to do, and how I just can't do them because I'm too tired. Sometimes I'll take a nap, hoping that it'll give me a boost, but it never really does. I feel like I'm just going through the motions, you know? I'm not really living, I'm just existing, and that's a really scary thought.

Table 12: Example 2 of pairwise comparison shown to annotators (Part 1). Underlined conversation is produced by higher-ranked simulator according to PSI-Bench.

Therapist: It sounds like that cycle of self-criticism feels almost automatic, almost like a default setting that pulls you into a heavier mood. When those thoughts surface, what sensations or images come with them—aside from the knot in your chest? How would you describe the space between noticing the thought and feeling the pull toward that "failure" narrative? **Patient:** I feel this tightness in my throat and my muscles just kind of... ache all over. In my mind I see a picture of myself at the desk, everyone staring, looking disappointed. There's a tiny gap—maybe a second—where I notice the thought, but then the feeling rushes in like a wave and pulls me straight into that "failure" story. My heart starts pounding and I get a cold sweat. It all happens so fast I barely have a moment to even breathe.	**Therapist:** It sounds like the mornings feel like a battle, with fatigue pulling you back even as you know you need to start the day. What small moments, if any, do you notice that bring a brief sense of relief or a spark of motivation? How does the feeling of "existing" rather than "living" show up in your thoughts or actions as the day goes on? **Patient:** I think it's the little things, like making a cup of coffee or taking a shower, that give me a tiny bit of energy and make me feel like I'm doing something, even if it's just small. Sometimes, if I'm lucky, I'll get a bit of momentum going and I'll manage to tackle a task or two, but it's always followed by a crash afterwards. I feel like I'm just putting on a mask, pretending to be okay, but deep down I'm still feeling exhausted and defeated. When I'm not working, I feel like I'm wasting my time, like I should be doing something productive, but I'm just too tired to even think about it. I guess it's like I'm trapped in this never-ending cycle of fatigue and guilt, and I don't know how to escape it.
Annotator 1 – B: "B seems most human and the realest of all the conversation in this block as far as I remember, having the best flow of conversation and seemingly something that could be actually heard when talking with a patient." **Annotator 2 – A**: "I chose Synthetic A because it actively incorporates the specific physical and sleep symptoms from the patient profile (e.g., tossing and turning, waking up late, muscle aches all over). It also uses more natural speech patterns, like hesitation and fragmented thoughts. Synthetic B is not terrible and correctly identifies the fatigue and work anxiety, but it feels a bit more generic. The patient in B talks in broad clichés (I am not really living, I am just existing, putting on a mask), whereas the patient in A grounds their anxiety in concrete physical sensations (tightness in the throat, muscle aches, pounding heart), which better matches the profile of someone suffering from severe unexplained physical exhaustion." **Annotator 3 – B**: "In B it´s not all negative, but a mix of hope and anxiety"	

Table 12: Example 2 of pairwise comparison (continued) and reasoning for annotators' non-tie choices.