Towards Scalable Therapist Training via Multi-Turn LLM-Based Client Simulation

Enea Gurra[1,*], Neha Deshpande[1]

[1]Technische Universität Berlin, Straße des 17. Juni 135, Berlin, 10623, Berlin, Germany

Abstract

Training therapists to provide interventions in highly sensitive psychotherapy domains—such as cases involving an individual's use of Child Sexual Abuse Material (CSAM)—is essential but presents unique challenges. These challenges include the limited availability of expert therapists to train new practitioners, and the resource-intensive nature of supervised role-play training, which commonly relies on expert therapists simulating client roles. Such approaches are difficult to scale, particularly given the global relevance of this domain and the need to support training across multiple languages. At the same time, the explicit nature of these conversations creates substantial technical, ethical, and deployment constraints for simulation-based training. This paper explores whether large language models (LLMs) can serve as realistic virtual clients to support scalable role-play-based training for psychotherapists, when guided by carefully designed prompts. We apply prompt-based simulation methods that use real therapy transcripts and case vignettes, without requiring model fine-tuning. Four LLMs of varying scales (9–123B parameters) were evaluated under zero-shot, few-shot, and retrieval-augmented few-shot prompting. Automatic measures were used to assess consistency and grounding, while expert therapists evaluated coherence, role consistency, realism, and clinical usefulness. Results indicate that both model scale and prompt design strongly affect simulation quality. The strongest performance was observed with the largest model combined with retrieval-augmented few-shot prompting, which produced simulated client responses rated as most clinically useful while maintaining high role consistency. These early findings suggest that carefully constrained LLMs can provide a scalable and promising support tool under appropriate safeguards for training therapists in complex, multi-turn interactions, even in highly sensitive domains of therapy.

Keywords

Large Language Models, Psychotherapy, Prompt Engineering, User Simulations, Therapist training

1. Introduction

Demand for mental health services is high worldwide, yet barriers to access remain significant [1, 2, 3]. At the same time, training clinicians and counsellors at scale is costly and logistically complex [4, 5]. These constraints can be even more pronounced for highly sensitive cases, where both the availability of appropriately trained supervisors and the feasibility of traditional role-play can be further restricted.

One such domain concerns prevention-oriented interventions for individuals reporting problematic sexual behaviour online, including cases involving Child Sexual Abuse Material (CSAM). Therapists working with such individuals require extensive preparation to navigate conversations, manage risk, and provide stigma-free interventions [6]. Reported cases of consumption of CSAM have increased sharply: the U.S. National Center for Missing & Exploited Children (NCMEC) documented an 87% rise in reports since 2019 [7]. Treating such cases is challenging, and opportunities for trainee therapists to practice interventions are limited. A long-standing approach to clinical training uses standardised patients (SPs)—trained actors or coached individuals who simulate clinical cases to support the development of communication and reasoning skills [5]. While effective, actor-based training is expensive, scheduling-intensive, and difficult to scale [4].

Recent advances in Large Language Models (LLMs) have demonstrated strong role-play capabilities and the ability to produce human-like conversations, making LLM-simulated clients a feasible alternative for practice scenarios [8, 9, 10, 11]. However, applying LLM simulation to CSAM-adjacent therapy training introduces constraints that differ from most prior work. Conversations about illegal sexual

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus

*Corresponding author.

e.gurra@campus.tu-berlin.de (E. Gurra); n.deshpande@tu-berlin.de (N. Deshpande)

behaviour (CSAM use) can trigger safety filters or generate undesirable outputs. Many mainstream LLMs are explicitly aligned to avoid such content, which hampers realistic simulation [12, 13], while unconstrained generation may produce harmful or inappropriate outputs. Additionally, privacy and governance requirements can impose on-premises deployment and restrict the use of third-party APIs. These domain-specific constraints shape both system design and evaluation.

Therapy client simulation in multi-turn interactions also raises methodological challenges independent of domain sensitivity. First, the simulated client must maintain a convincing and internally consistent persona across turns. Real clients present distinctive personal histories, emotional states, and conversational quirks that unfold over time. Achieving such nuance is difficult: without careful prompting, LLMs tend to produce generic responses, drift out of character, or forget key backstory details over longer interactions [14, 15]. Nevertheless, evaluation is inherently complex. Unlike task-oriented systems, a simulated client cannot be measured by simple success/failure criteria. The quality of a model's replies depends on whether they are plausible for the client to say at that point in the conversation—i.e., they fit the meaning of the exchange, the emotional tone, and the assigned persona. Standard overlap-based metrics often correlate poorly with human judgments in open-domain dialogue [16]. A robust evaluation framework must therefore combine quantitative measures with expert human assessment.

This paper, therefore, presents a case study of adapting prompt-based LLM simulation to a safety-critical, high-constraint therapy training context. It investigates the use of an LLM-based chatbot that simulates client behaviour in therapist–client interactions, addressing CSAM consumption, aiming to create a controllable practice environment. The present study focuses on the client side of multi-turn therapy interactions and examines how prompting strategies and model choice influence simulation quality and the potential usefulness of such simulations for clinical training. Such client personas allow therapists to rehearse assessment and intervention skills without the recurring costs and logistics associated with live actors. This study is situated within the EU-funded STOP-CSAM[1] program, which offered appointment-based, real-time multilingual text chat with qualified therapists. Experiences from this clinical context underscored a shortage of training resources for therapists working with individuals presenting problematic sexual behaviours. Throughout this paper, the term "client" refers to the simulated help-seeker/patient. An example interaction between a human therapist and an LLM-simulated client is shown in Figure 1.

In summary, this paper makes the following contributions:

- Documents practical prompt-based techniques that elicit stable, client-only behaviour from off-the-shelf LLMs and supports grounding in realistic case materials.
- Reports an end-to-end setup for multi-turn therapist–client interactions, including a lightweight web interface, a transcript/logging pipeline, and a mixed-methods evaluation protocol that combines automatic, reference-based metrics with expert clinician ratings of realism, coherence, role consistency, and clinical usefulness.
- Benchmarks four LLMs spanning 9B–123B parameters under three prompting strategies (zero-shot, few-shot, and retrieval-augmented few-shot) using automatic metrics.
- Reports a focused expert study with live sessions and distils practical findings on how model scale and prompting affect realism, role consistency, and perceived clinical usefulness
- Documents a privacy-preserving deployment pattern for sensitive scenarios (e.g., CSAM-related training cases) and outlines ethical safeguards and reporting practices for responsible use.

[1] https://stop-csam.charite.de/en/

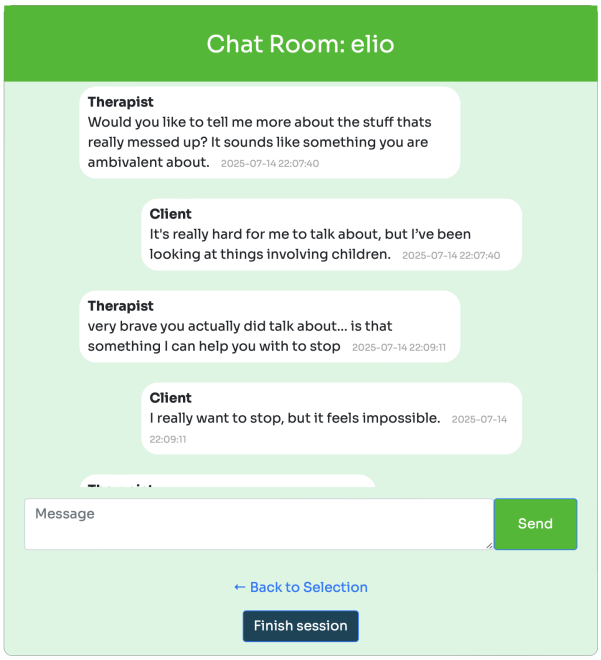


Figure 1: Screenshot of the prototype training interface showing an example exchange between a therapist and an LLM-simulated client.

2. Related Work

2.1. LLMs for Clinical Simulation and Training

Artificial Intelligence (AI) is increasingly used in mental health support and training, from automated mental health apps to training tools for clinicians. Systems such as Woebot[2] leverage CBT techniques and empathetic language; a randomised trial reported mood improvements among college students [17]. Simulated patients (SPs) are highly valuable in training psychologists and psychiatrists, but actor-based programmes are costly and limit scalability [4] and can impose emotional strain on SPs [18, 19]. LLMs offer scalable alternatives: Chen et al. [8] report feasibility of psychiatrist–patient simulations with expert ratings; Wang et al. [11] introduce CBT-grounded patient models (PATIENT-Ψ), with trainees and experts rating them favorably against baselines; and Li et al. [20] present a model-agnostic dialogue-and-feedback framework with more authentic flows than generic chatbots. Fully synthetic corpora have also emerged: Duro et al. [21] generate multilingual counselling dialogues that reproduce several human-like patterns, while Lee et al. [22] release CACTUS, a CBT-structured dataset with diverse personas and improved counselling-skill ratings for models trained on it. Collectively, these studies indicate that LLMs can approximate aspects of client language and affect, while underscoring the need for careful clinical validation and safety oversight [23, 24].

2.2. Prompting strategies for simulated clients

In the context of generative AI, a prompt refers to the input provided to a model with the purpose of shaping or directing the content it generates [25]. LLMs have shown a remarkable ability to adopt fictional personas and engage in role-specific dialogues. In psychiatric and counselling contexts, role-play prompts can elicit patient-like dialogue that experts judge as realistic [8]. Persona conditioning further stabilises style and behaviour across turns [26]. Domain-informed prompting can improve fidelity: Louie et al. [9] combine expert-derived principles with a principle-adherence procedure, yielding sizable gains in response quality and rule following. Prompt-only methods retain model weights; behaviours are induced via instructions and a few exemplars [27]. Njifenjou et al. [28] describe

[2] https://woebothealth.com

role-play zero-shot prompting as an efficient, cost-effective way to customise open-domain conversation, showing that a well-crafted prompt with an instruction following model can match or even surpass fine-tuned dialogue models on human evaluations in French tasks. In parallel, Brown et al. [27] demonstrate that LLMs are few-shot learners that can perform tasks from prompts without gradient updates when given a handful of examples and instructions. These findings show that in-context learning can be a practical alternative to task-specific fine-tuning for different applications.

2.3. Evaluating Simulated Clinical Dialogue

Evaluating open-ended dialogues, especially in a therapeutic context, is notoriously complex. Unlike tasks with a single correct answer, a conversation can be valid in many ways, and quality has multiple facets (e.g., coherence, relevance, empathy, fidelity to persona).

Traditional n-gram overlap metrics such as BLEU [29] and ROUGE [30] correlate weakly with human judgments in open-ended dialogue, largely because many different responses can be appropriate for a given context, so exact word overlap is a poor proxy for quality [16, 31]. Recent learned or embedding-based metrics aim to capture meaning rather than surface form: for example, BERTScore compares contextualised embeddings, and BLEURT uses a learned regression objective, both of which can better reflect semantic adequacy when reference responses are available [32, 33].

Consequently, mixed-methods designs are prevalent: Qiu and Lan [10] combine semantic consistency with expert choices among candidate replies; Njifenjou et al. [28] report automatic scores as auxiliaries and centre human ratings (coherence, engagingness, humanness); Hsu et al. [34] evaluate practice-and-feedback systems with surveys, interviews, and usage logs alongside standard metrics; Louie et al. [9] and Bodonhelyi et al. [35] rely on expert judgments of authenticity and preparedness. This literature supports an evaluation strategy that pairs reference-based semantics with expert therapist ratings of realism, coherence, persona fidelity, and usefulness.

3. Methods

3.1. Data Collection and Preprocessing

For this study, data from the domain of CSAM prevention was required. Data was obtained from a prior online CSAM prevention programme that offered scheduled, real-time text-chat counselling with qualified therapists. The dataset included two realistic client vignettes (Elio and Marco) and counselling transcripts comprising (i) 29 therapist-training role-plays, where experienced therapists acted as clients to train clinicians-in-training, and (ii) authentic sessions with two clients across a total of seven sessions (two and five sessions, respectively). The vignettes represent fictitious but clinically realistic cases, developed by experienced psychotherapists drawing on their professional experience with clients in this domain.

These vignettes were converted from PDF to structured JSON persona profiles via template-guided extraction with manual checks. Transcripts were standardised by removing programme-context statements, collapsing rapid successive client messages, enforcing one message per turn, and retaining only text and essential turn-level metadata. Non-English and incomplete role-plays were excluded. After pre-processing, 20 of 29 conversations were retained (on average, 15 client turns and 20 minutes per dialogue). The remaining 9 were incomplete, test dialogues, or non-English. Two conversations were based explicitly on the vignette “Marco” and one on “Elio”.

3.2. Prompt Design

This work mainly leverages prompt engineering to simulate realistic clients seeking therapy from instruction-tuned LLMs across multi-turn dialogues. Three prompting strategies were selected, iteratively refined, and compared: a structured zero-shot prompt, a static few-shot prompt, and a retrieval-augmented few-shot prompt (FSR) that adds turn-wise exemplars at inference time. Every

prompt design uses the same structured scaffold with role-play prompting. Prompt variants were iteratively refined based on automatic metrics and qualitative checks. Full prompt texts and templates are in Appendix A.

Prior studies suggest that structuring prompts enhances comprehension and response consistency [25, 36]. Building on these findings, all prompts share and build upon the following template:

1. **System instructions** Instruct the model to act solely as the client (never as therapist, narrator, or assistant).
2. **Guidelines for client's utterances** Consists of a checklist that sets the dialogue language, the limit turn length, and tone, and reminds the model to remain in character throughout the exchange.
3. **Client Persona** Injects the relevant persona profile (Elio/Marco) to ensure consistent, persona-aligned responses.
4. **Summary of earlier trimmed conversation** When the dialogue approaches the model's context-window limit, earlier turns are truncated and summarised into a brief recap that replaces the removed text to preserve coherence under a sliding-window regime. We use an LLM-generated summary capped at around 100 tokens.
5. **Conversation History** Therapist and client utterances are appended turn by turn, every run dynamically.
6. **Consistency Reminder** Reiterates that the client must stay in character and keep the responses concise.

Role-play prompting instructs an LLM to adopt a specific persona, goals, and constraints so that responses consistently reflect a defined role and communication style across turns [37]. Role-play was used in combination with all other strategies tested.

3.2.1. Zero-shot Prompting

In zero-shot prompting, the LLM is guided solely by a detailed instruction-based prompt without explicit examples of desired outputs [27, 38]. The zero-shot variant comprised only the shared scaffold, relying on instruction following without demonstrations. It served as a baseline to assess how well models could realise a client persona from instructions alone. Figure 5 in Appendix A shows the zero-shot prompt used to simulate the role of the client.

3.2.2. Static Few-shot Prompting

For each persona, short exemplars (1–3 therapist–client exchanges) were appended to illustrate opening, disclosure, and late-session behaviours. Exemplars match the dialogue schema and vary in length to avoid anchoring. To minimise leakage, one human-to-human transcript (Marco) was reserved exclusively for exemplars and excluded from evaluation. For Elio, a Czech transcript was machine-translated for exemplar crafting without verbatim reuse from the English evaluation transcript. The Examples section appears after context and before the closing reminder. Figures 6 and 7 in Appendix A show this section for Elio and Marco.

3.2.3. Retrieval-Augmented Few-shot Prompting

The third strategy incorporates a dynamic retrieval mechanism that supplies the model with relevant examples at each turn, aiming to mimic how a client speaks in similar cases. At each turn, the latest therapist input was embedded with the `intfloat/e5-large-v2` sentence-embedding model [39, 40] and used to query a FAISS index [41] built over curated therapist-client snippets from the

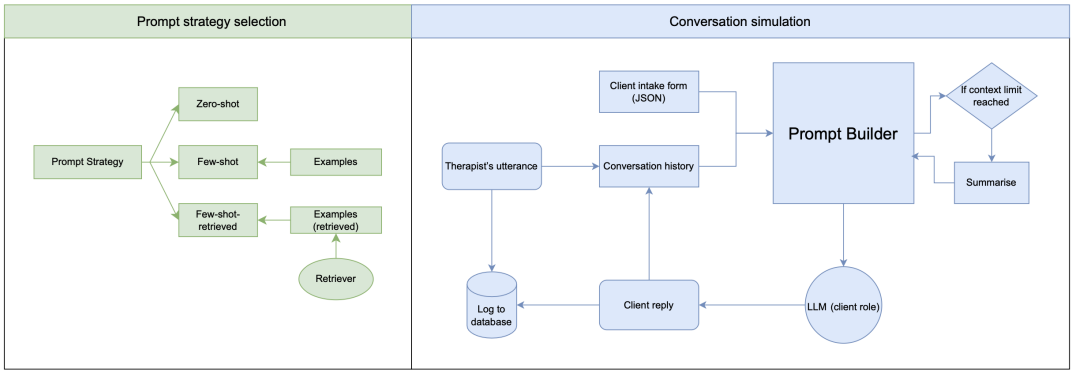


Figure 2: Client-simulation pipeline: First, the strategy is selected (zero-shot or few-shot, or few-shot-retrieved), and in accordance examples are retrieved. The Prompt Builder composes the system prompt conditioned on the conversation history. The LLM generates client replies, and all turns are logged for evaluation.

normalised transcripts. The top-10 retrieved candidates are then re-ranked with a cross-encoder (`cross-encoder/stsb-roberta-base`) [42, 43]. The example block was added in the prompt as shown in Figure 8 in Appendix A. In practice, we use K = 10 exemplars per turn. Hereafter, we refer to this strategy as few-shot-retrieved (FSR).

3.3. Model Selection

Candidate models were screened for (i) size and latency for real-time latency constraints in a live chat setting, (ii) dialogue quality, (iii) robustness in role-playing the client, and (iv) absence of hard filters on CSAM-related vocabulary. Common screening failures were role reversal, character breaks, policy-driven refusals after several turns (refuses to discuss CSAM-related material), and multi-turn over-generation in a single reply.

Four models spanning 9B–123B parameters were selected: Gemma-2-9B (base; `gemma2:9b`[3]), Mistral-Small-Instruct-2409 (22B; `mistral-small:22b`[4]), Qwen2.5-72B-Instruct(72B; `qwen2.5:72b`[5]), and Mistral-Large-Instruct-2411 (123B; `mistral-large:123b`[6]) [44, 45, 46, 47]. This set enabled analysis of whether scale and architecture affected simulated-client quality while remaining deployable on-premises. The ollama tag for each model used is shown in parentheses.

For brevity, throughout the remainder of this paper, Mistral-Small-Instruct-2409 will be referred to as "Mistral Small", Qwen-2.5-72B-Instruct as "Qwen 2.5", and Mistral-Large-Instruct-2411 as "Mistral Large". "Gemma 2" remains unchanged.

Figure 2 shows the pipeline of the client-simulation methodology.

4. Experiments and Evaluation

We adopt a two-stage evaluation strategy. Experiment 1 (section 4.2) provides a controlled, repeatable screening setup that enables systematic comparison of models and prompting strategies across identical therapist turns. Experiment 2 (section 4.3) complements this by assessing interactive performance and training relevance via expert therapist judgments in live sessions. Together, these experiments triangulate consistency/grounding signals with clinician-perceived realism and clinical usefulness.

[3] https://ollama.com/library/gemma2:9b
[4] https://ollama.com/library/mistral-small:22b
[5] https://ollama.com/library/qwen2.5:72b
[6] https://ollama.com/library/mistral-large:123b

4.1. Experimental Setup

All experiments were executed on an institutional server (Ubuntu 22.04.4 LTS, Python 3.12.9, CUDA 12.4). To avoid transmitting sensitive prompts off-site, all LLMs were hosted locally via Ollama; end-to-end latency per client turn ranged from under 1 s (Gemma 2) to about 9 s (Mistral Large). Data ingestion, pre-processing, and metric computation were scripted in Python. Live sessions were conducted through a password-protected Flask web application; all messages were timestamped and persisted with model, prompt, and session metadata.

4.2. Experiment 1: Automatic Screening Evaluation on Fixed Transcript Turns

4.2.1. Design and Procedure

The first experiment used a set of real therapy session transcripts to evaluate simulated client responses in a controlled manner. Two complete human-human, role-play therapy transcripts (personas: Marco, Elio) were used to fix the therapist side of the dialogue. For each run, the therapist's utterances were replayed verbatim, and the model generated the client's reply one turn at a time. Fixing therapist turns controls for therapist variability and isolates the effect of model choice and prompting strategy, enabling fair comparisons under identical conversational stimuli. This transcript-replay setup yields a controlled, reference-based screening protocol in which generated client utterances were compared with the corresponding human client utterances from the source transcript.

Three prompting strategies were crossed with four models and two personas, producing 24 simulated dialogues for automatic scoring. It is important to acknowledge that, in psychotherapy, no single "correct" client response exists for a given therapist prompt. Therefore, differences from the transcript should not necessarily be regarded as errors. Instead, automated scores were treated as screening signals and are interpreted cautiously and complemented by expert human judgements in Experiment 2.

4.2.2. Evaluation: Automated Metrics

Automatic metrics were computed as screening signals by comparing each model-generated dialogue with the corresponding real session transcript (Marco or Elio), both turn by turn and over the dialogue as a whole. Turn-level similarity to the reference client utterance was computed using ROUGE-1/L [30] and BERTScore F1 [32]. Sentence-BERT [42] cosine similarity provided an additional turn-level semantic signal. At the session level, Distinct-1/2 quantified repetition [48], and average words per client turn captured verbosity control. Prior studies have noted that response length has a strong effect on quality: too short can seem uninterested or evasive, while too long can feel like the bot is rambling [49]. Tracking average length helps ensure each configuration's verbosity remains within a plausible human range.

Persona adherence was assessed by converting each intake persona into a short first-person narrative and computing BERTScore recall/precision between the concatenated client turns and that narrative. The narrative was generated by prompting Mistral Large to rewrite the structured JSON persona into a concise first-person self-description. In this setup, recall estimates how much persona information made it into the dialogue (coverage), while precision provides a proxy for persona-groundedness, penalising content that is weakly supported by the persona narrative (lack of extraneous or contradictory content). Finally, as an overall measure of session-level similarity to the reference transcript, SBERT cosine similarity between the concatenated sequence of generated client turns and the concatenated sequence of ground-truth client turns was computed, treating each session as two long texts. All automatic metrics were interpreted alongside qualitative, human-centred judgements, given the multiplicity of plausible client responses in therapy dialogue.

4.3. Experiment 2: Live Therapist-LLM Interaction

4.3.1. Design and Procedure

Experiment 2 complements the transcript-based screening in Experiment 1 by evaluating the simulated clients in an interactive setting with expert therapists. Two licensed therapists conducted live counselling sessions with the simulated clients via a web interface and evaluated the sessions. One clinician had roughly 15 years of specialised experience treating clients who report consuming CSAM; the other was an early-career clinician recently trained in this area. These two profiles provide perspectives reflecting both extensive domain expertise and a plausible end-user group for training-oriented deployments.

Given the sensitivity of CSAM-related content and the risk of misinterpretation without appropriate clinical context, participation was restricted to licensed clinicians with relevant experience; students and non-specialists were not eligible because of safety and validity. Given the small and specialist sample, the study is exploratory and analyses are descriptive and trend-oriented. To make efficient use of limited expert time while still enabling comparisons across prompting strategies, Experiment 2 focused on the smallest and largest models in our pool (Gemma 2 and Mistral Large) and a single standardised client persona (Elio). This provides a high-contrast comparison across model capacity while keeping the number of live sessions tractable. Each therapist completed controlled assessment sessions individually and did not communicate with one another.

After logging in, therapists selected a prompting strategy (zero-shot, few-shot, or retrieval-augmented few-shot), a model (Gemma 2 or Mistral Large), and the client persona. Sessions were conducted in real time with a minimum of 15 turns to encourage comparable depth and to reduce early role breaks. The system logged all content and metadata, producing complete transcripts for analysis. After each session, the therapist completed a brief questionnaire embedded in the interface to rate the quality of the simulated client along several dimensions.

4.3.2. Evaluation via Human Judgements

Immediately post-session, therapists completed a 20-item questionnaire covering four constructs: *Realism* (5 items), *Coherence* (7), *Role Consistency* (6), and *Clinical Usefulness*. *Realism* captures whether the simulated client presents as a plausible person in context, including whether autobiographical details are provided in a consistent and believable manner; *Coherence* reflects local and global consistency of responses across turns; *Role Consistency* assesses adherence to the assigned client persona without therapist-role reversal or character breaks; and *Clinical Usefulness* reflects whether the simulated client supports meaningful therapist skills practice in a training context (e.g., practising engagement, questioning, reflections, and managing avoidance), rather than clinical decision-making. This was informed by evaluation frameworks from prior dialogue-system research and virtual-patient studies, as well as standard criteria for coherent, human-like conversation [28, 22, 35, 49, 50, 9]. Responses were recorded on five-point Likert scales (1 = Strongly disagree to 5 = Strongly agree). Full instrument and item provenance can be found in Appendix B.1.

No counterbalancing or order manipulation was applied; session order followed therapist availability and scheduling constraints. Given the exploratory nature of the study and the limited number of raters and sessions, we report descriptive summaries and exact small-sample tests to characterise trends. For transparency, effect sizes, exact p-values, and agreement estimates are included. More details on the expert evaluation, including the questionnaire, web interface used, details on the statistical reports and analysis are documented in Appendix B.

5. Results

This section reports the results of automatic metrics (Experiment 1; Section 5.1) and then summarises clinician ratings with descriptive statistics and exact small-sample tests (Experiment 2; Section 5.2).

5.1. Automatic Evaluation

Table 1 provides an overview of key metrics, averaged across the two clients, while more detailed per-client breakdowns are given in Table 5 in Appendix C. Then, overall performance across models and prompts, turn-level similarity trends, session-level similarity, verbosity, and persona consistency are examined.

Table 1
Automatic evaluation (Experiment 1) by model and prompting strategy (zero-shot, few-shot, few-shot-retrieved/FSR), averaged across the two client personas. ROUGE-1/L measures lexical overlap; BERTScore (F1) and SBERT measure turn-level semantic similarity; Dist-1/2 measures response diversity; Words/turn is the mean word count per client turn; Persona Precision/Recall quantify persona adherence; SBERT session is session-level embedding similarity. Bold marks the best prompting strategy within each model for each metric (higher is better; for Words/turn, bold indicates highest verbosity).

Model	Prompt	ROUGE-1	ROUGE-L	BERTScore (F1)	SBERT	Dist-1	Dist-2	Words/turn	Pers. Prec	Pers. Rec	SBERT session
gemma2:9b	few-shot	0.109	0.093	0.504	0.151	0.413	0.716	13.22	0.496	0.497	0.383
	few-shot-retrieved	**0.147**	**0.119**	**0.535**	**0.228**	**0.487**	0.803	10.26	**0.507**	0.514	**0.682**
	zero-shot	0.130	0.102	0.515	0.223	0.470	**0.818**	**13.46**	0.500	**0.523**	0.667
mistral-large:123b	few-shot	0.190	0.154	0.547	0.314	0.474	**0.879**	24.00	0.508	**0.554**	0.783
	few-shot-retrieved	**0.331**	**0.305**	**0.603**	**0.423**	**0.493**	0.873	18.18	**0.510**	0.549	**0.813**
	zero-shot	0.186	0.146	0.535	0.277	0.448	0.875	**27.79**	0.494	0.553	0.707
mistral-small:22b	few-shot	0.131	0.110	0.518	0.177	0.429	0.735	14.02	0.538	0.539	0.577
	few-shot-retrieved	**0.228**	**0.202**	**0.565**	**0.329**	0.511	0.832	12.86	**0.541**	0.564	0.685
	zero-shot	0.150	0.120	0.527	0.235	**0.517**	**0.875**	**15.83**	0.540	**0.565**	**0.746**
qwen2.5:72b	few-shot	0.223	0.171	0.541	0.295	0.358	0.714	32.39	0.489	0.531	**0.803**
	few-shot-retrieved	**0.228**	**0.180**	**0.557**	**0.340**	0.388	0.783	30.25	**0.499**	**0.556**	0.760
	zero-shot	0.191	0.140	0.533	0.296	**0.397**	**0.809**	**33.59**	0.484	0.538	0.737

5.1.1. Overall Performance across Models and Prompts

Performance varied by model-prompt combination. In general, larger models achieved higher similarity to the human reference dialogues, with Mistral Large (123B) in FSR yielding the best aggregate scores. Qwen2.5 (72B) performed strongly (its best session-level SBERT near 0.80 with few-shot), whereas Gemma 2 (9B) underperformed, particularly with few-shot exemplars. For the smaller models, zero-shot exceeded few-shot, whereas (FSR) partially recovered performance.

5.1.2. Turn-level Similarity Dynamics

Figure 3 shows how closely each model's responses matched the reference client utterances at each turn, as measured by BERTScore (F1). The similarity trajectories across turns fluctuated. Pronounced dips occurred when the reference client answers were very short (e.g., 'yes', 'okay') or when the therapist produced a statement without a follow-up question. These dips often reflect metric sensitivity to short utterances rather than genuine failures. On substantive turns, larger models more frequently matched the reference semantics, and FSR tended to stabilise early- and late-turn similarity. Providing relevant exemplar dialogues helped the larger models sustain more stable semantic similarity throughout the session.

5.1.3. Verbosity and diversity

Figure 4 shows the number of words per simulated client response at each turn. FSR consistently reduced verbosity relative to zero-shot (e.g., Mistral Large (123B): around 28 to 18 words/turn), suggesting that gains in similarity under FSR are not merely length effects. All models maintained high lexical diversity (Distinct-2 typically >0.80), indicating no collapse into generic responses.

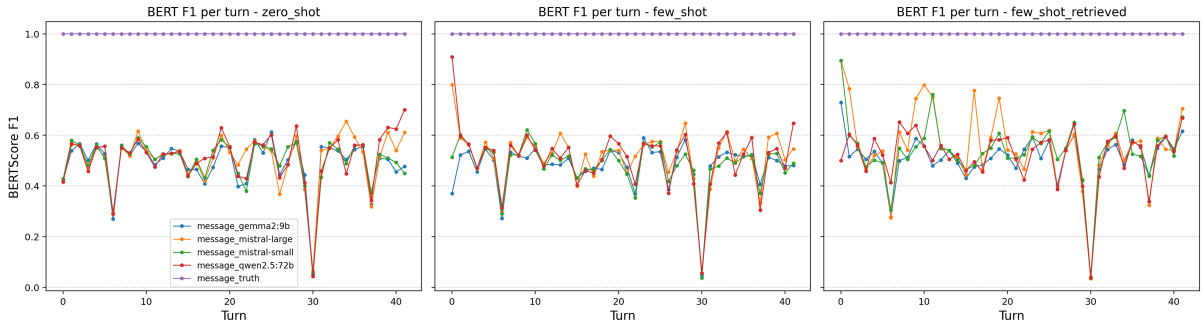


Figure 3: BERTScore (F1) scores/values between the model reply and the reference client utterance, per turn, for each model and prompt condition.

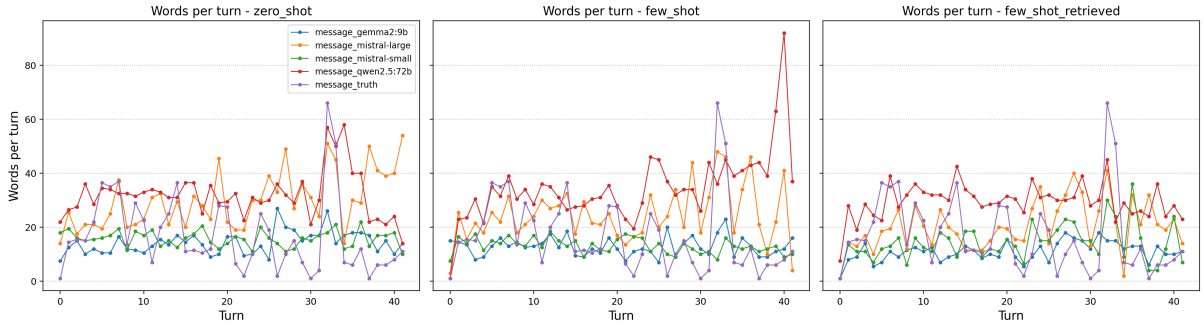


Figure 4: Words per simulated-client turn across the session, by model and prompt condition.

5.1.4. Session-level Similarity

Mistral Large (123B) achieved the strongest whole-session resemblance when paired with the FSR prompt. Qwen 2.5 (72B) benefited most from static few-shot, with retrieval offering no additional gain. For smaller models, Gemma 2 (9B) and Mistral Small (22B), zero-shot remained the most reliable baseline, and few-shot sometimes degraded similarity. Overall, targeted exemplars help higher-capacity models, whereas lighter models are more stable without them.

5.1.5. Persona Adherence

A crucial aspect of therapy-client simulation is consistency with the client's backstory and profile (persona). Table 1 summarise the results for both client personas. Overall, all models stayed reasonably true to their assigned character. Persona precision hovered around 0.50–0.57 and recall benefited from FSR without notable precision trade-offs. Smaller models sometimes showed slightly higher precision (stricter adherence), while larger models added elaborations yet retained acceptable precision.

5.2. Expert Evaluation

5.2.1. Exploratory Trend Summary

Expert ratings are interpreted as indicators of whether the simulated client can support supervised therapist skills practice in CSAM-related prevention therapy (i.e., training utility), rather than as evidence of clinical effectiveness. Table 2 summarises medians and interquartile ranges (IQR) for each model–prompt configuration. One therapist completed all six model–prompt combinations (zero-shot, few-shot, FSR for each model), while the other completed four combinations, omitting Gemma 2 with few-shot and few-shot-retrieved. Given the small, specialist sample and partially unbalanced coverage of conditions (11 total across two therapists), results are treated as exploratory and descriptive.

Across the rated dimensions, Mistral Large (123B) generally showed higher central tendencies than Gemma 2 (9B). Under FSR, Clinical Usefulness for Mistral Large reached a median of 5.00 (IQR 0.00) with Role Consistency near ceiling (4.75–5.00). Reflecting on this configuration, one clinician remarked, "It's really amazing"; the only concern raised was a remaining interaction-format limitation of the current interface (single response per turn): "there is always exactly one response—never none and never more than one. In a human chat, the number of responses varies from 0 to x." Gemma 2 scored highest in zero-shot but declined under few-shot and FSR on realism and coherence, while maintaining high role consistency. There were isolated settings where Gemma 2 was competitive; for example, Gemma 2 in zero-shot exceeded the larger model in few-shot on some dimensions. However, the dominant pattern is that both model capacity and prompting strategy materially shaped perceived quality and training value. Overall, the expert ratings suggest that (i) larger models provide more training-relevant interactions, and (ii) retrieval-augmented exemplars can improve training utility for high-capacity models, whereas static few-shot prompts may induce failure modes that reduce perceived quality—especially for smaller models. Given the small, expertise-focused design, these results are treated as exploratory and illustrative of expert judgement trends.

Table 2
Expert evaluation results from live sessions: medians with interquartile ranges (IQR) in parentheses and aggregate quality scores. The study involved two licensed therapists (n=2) and 11 total sessions. One therapist completed all six model–prompt combinations (zero-shot, few-shot, and FSR for each model), while the other completed four combinations, omitting Gemma 2 with few-shot and few-shot-retrieved. Bold values indicate the best score within each model block (across prompts) for each dimension and for the overall score. "Overall score" is the median of the composite quality score aggregated from the four dimensions for each session.

		Realism	**Coherence**	**Role Consistency**	**Usefulness**	**Overall score**
Model	**Prompt**	Median (IQR)	Median (IQR)	Median (IQR)	Median (IQR)	Median
mistral-large	zero-shot	**4.70** (0.30)	**4.36** (0.64)	**5.00** (0.00)	4.75 (0.25)	**4.72**
	few-shot-retrieved	4.50 (0.30)	4.29 (0.14)	4.83 (0.17)	**5.00** (0.00)	4.66
	few-shot	3.60 (0.40)	3.79 (1.07)	4.75 (0.25)	3.50 (1.50)	3.70
gemma2	zero-shot	**3.75** (0.35)	**3.07** (0.93)	**4.88** (0.04)	**3.50** (1.50)	**3.62**
	few-shot-retrieved	2.60 (0.00)	2.43 (0.00)	4.50 (0.00)	2.00 (0.00)	2.52
	few-shot	2.80 (0.00)	2.14 (0.00)	4.50 (0.00)	1.00 (0.00)	2.47

5.2.2. Effect of Prompt Strategy on Rated Quality

Prompt effects differed by model capacity. For Mistral Large, static few-shot underperformed relative to zero-shot, while adding retrieval (FSR) restored or exceeded performance on several dimensions, most notably Clinical Usefulness. For Gemma 2, exemplar-based prompting (few-shot and FSR) tended to reduce Realism and Coherence relative to zero-shot. This interaction suggests that retrieval-augmented exemplars can be beneficial for higher-capacity models in this setting, whereas exemplars may induce undesirable response patterns in smaller models under the same constraints.

5.2.3. Statistical tests

Due to the small number of raters and limited matched pairs, inferential statistics are underpowered and are reported primarily for transparency rather than definitive inference. In the zero-shot condition, paired contrasts favoured Mistral Large over Gemma 2 across all dimensions (median differences: Realism $+0.95$, Coherence $+1.29$, Role Consistency $+0.13$, Usefulness $+1.25$), noting the very small number of pairs (see Table 8). Within the larger model, Friedman ranks for Realism suggested zero-shot $>$ FSR $>$ few-shot. All exact test outputs, p-values, agreement estimates, and Bland–Altman summaries are reported in Appendix C for completeness and should be interpreted descriptively.

5.2.4. Engagement and Qualitative Notes

Table 3 reports the number of utterances per session and the average client turn length from the human evaluation dialogues. Sessions with Mistral Large often lasted longer and were more talkative. Under the FSR prompt, Mistral Large averaged about 86 client–therapist turns per session, compared with roughly 42 turns for Gemma 2 under the same prompt.

Qualitative feedback helped contextualise these engagement differences; the following comments are illustrative of themes that emerged in the open-ended post-session responses. For Mistral Large under static few-shot, therapists reported a repetitive stalling pattern, with one clinician noting that "at a certain point every response was 'yes..., but...'". In Mistral Large zero-shot sessions, one therapist observed that the simulated client could feel "a bit too emotionally reflective" and "too hesitant to commit to behavioural change", repeatedly returning to fears and "what if" hypotheses. In contrast, under FSR, the simulated client more often produced specific, on-topic elaborations that supported continued exploration. No explicit client-role breaks were reported by raters in these sessions.

Table 3
Engagement characteristics in the live expert study: mean number of utterances per session and mean client words per turn, aggregated across completed human-evaluation sessions for each model–prompt condition (n=2 therapists; 11 sessions total; partially unbalanced across conditions).

Model	Prompt	# Utterances	Avg. words/turn (client)
gemma2:9b	few-shot	40.0	12.85
	few-shot-retrieved	42.0	15.14
	zero-shot	64.5	15.04
mistral-large	few-shot	53.0	29.56
	few-shot-retrieved	86.0	19.91
	zero-shot	54.0	30.12

6. Discussion

This work examines the feasibility of prompt-guided LLM-based client simulation as a support tool for supervised therapist skills practice in CSAM-related prevention therapy, focusing on multi-turn role-play. We triangulated a controlled transcript-replay screening experiment with a live expert interaction study, interpreting outcomes in terms of perceived coherence, role consistency, plausibility (realism), and training utility rather than clinical effectiveness. The results highlight how model capacity and prompt design jointly shape training-relevant behaviour under sensitive-domain constraints.

The size of the model emerged as the strongest driver of performance. The Mistral Large model generally produced responses that the therapists rated more favourably on realism, coherence, and perceived usefulness than the Gemma 2 model. Its utterances were more nuanced and context-aware, whereas the smaller model was prone to repetition and occasional missed questions. These observations are consistent with prior findings on scale advantages in fluency and coherence [27, 51]. The trade-off is computational: achieving interactive speed with the largest model required specialised hardware, while the smallest model was responsive but less convincing. Mid-sized models such as Mistral Small (22B) or Qwen 2.5 (72B) sometimes approached large-model quality on automatic scores, suggesting that efficiency techniques or fine-tuning could make smaller systems viable in constrained settings.

The prompt shaped outcomes, but its effect depended on model capacity. For weaker models, simple zero-shot prompts were more reliable than static few-shot examples, which sometimes diluted focus and lowered both automatic and expert ratings. By contrast, the largest model benefited from contextually retrieved examples (retrieval-augmented few-shot), which anchored responses in case-relevant material and strengthened persona fidelity. In short, exemplar design interacts with capacity: poorly matched exemplars can bias smaller models, whereas case-aware retrieval stabilises behaviour in stronger models.

Sustained adherence to the client's role is a key requirement in training scenarios. Across conditions, models remained largely in role, and systematic breaks were not reported. Larger models occasionally improvised additional details, but these additions were typically reasonable rather than contradictory. A structured intake scaffold and periodic role reminders likely supported this stability, echoing evidence that role-play prompting improves persona fidelity [37, 9]. Minor slips occurred, but rarely disrupted conversations.

Finally, automatic metrics were useful for screening settings, but did not fully capture conversational quality. They aligned with broad trends (e.g., large outperforming small, benefits from FSR) but diverged in cases such as Gemma 2 with FSR, where lexical overlap inflated scores despite the rating of the interactions by the therapists as generic or repetitive. This underscores the limits of reference-based evaluation in open-ended dialogue [16]. Metrics such as Distinct-1/2 and persona adherence provided complementary signals, but clinical usefulness ultimately required expert judgement. A practical workflow is therefore to use automatic metrics for pre-screening, followed by targeted human evaluation of promising conditions.

The findings support several deployment guidelines for training-focused use. First, matching prompting to model capacity: smaller models are more reliable under zero-shot prompting, whereas larger models benefit from few-shot-retrieved prompting that grounds responses in case-relevant material and strengthens persona fidelity. Second, using prompts that explicitly reinforce the client role—such as intake scaffolds and periodic role reminders—to reduce drift and derailment. Third, applying light verbosity control so that turn lengths remain appropriate for trainee pacing without sacrificing fidelity. Fourth, expecting longer sessions under larger models with FSR; training setups should plan for time-boxed closures and manage fatigue accordingly. Lastly, the evaluation should combine automatic screening with targeted expert review. These insights inform how LLMs might be developed into credible and useful tools for therapy training.

The findings suggest a practical training use case: structured role-play exercises in which trainees practise first-session micro-skills (e.g., alliance-building, reflective listening, risk-sensitive questioning, responding to avoidance or minimisation) with subsequent supervisor debrief. In such a workflow, the simulator is best treated as a practice partner that supports repetition and standardisation, not as a substitute for supervision or as clinical decision support. Several expert comments further pointed to interactional aspects that matter for training realism (e.g., the perceived naturalness of response timing and conversational pacing), suggesting that interface design choices are part of training validity, not just model performance.

7. Limitations

While our study provides promising evidence for LLM-based client simulation, several limitations should be acknowledged. The expert evaluation was necessarily limited by the domain's specialised nature and safety-critical context. We therefore restricted participation to licensed clinicians with relevant experience and did not recruit students or non-specialist therapists. Two clinicians completed 11 sessions, which supports credible, safe interaction but limits statistical power and generalisability; we accordingly emphasise descriptive trends and report full small-sample statistics and agreement estimates for transparency. More broadly, each configuration should be interpreted as a small number of case studies rather than a definitive benchmark across therapists. The findings are thus useful for identifying trends, but a larger study with more therapists and more varied cases is needed to statistically confirm the results.

Experiment 1 uses transcript replay as a controlled screening protocol rather than a notion of "correct" client responses: in psychotherapy multiple client utterances may be plausible for the same therapist prompt. Reference-based automatic metrics therefore, provide only indirect signals (e.g., grounding or consistency) and must be interpreted cautiously alongside expert judgment.

For ethical and data-protection reasons, all experiments were conducted on a university server without transmitting content externally. This setup ensured that sensitive CSAM-related prompts and

transcripts remained on-premises, but it also meant that widely used larger models could not be utilised. As a result, the comparative baseline is limited to self-hostable, open-weight, or research-licensed models. External validity may therefore be affected, as the absolute ceiling of achievable quality was determined by the best local model available under governance constraints.

Finally, although a set of role-play conversations was obtained from the STOP-CSAM project, its volume and diversity were insufficient to support robust fine-tuning of larger models. Scarcity of domain-specific data in this area remains a central challenge and is further constrained by ethical and legal restrictions on collection and sharing. In addition, only two client profiles were available, which reduced persona diversity and limited the generalisability of the findings.

8. Ethical Considerations

This work concerns a highly sensitive clinical training context (CSAM-related prevention therapy) with clear legal and ethical implications. The system is intended solely as a clinician-facing training and supervision aid. It is not a public-facing chatbot, not accessible to people seeking CSAM and is not intended to provide care, diagnosis, or crisis support.

Participation was restricted to licensed clinicians with relevant experience. Participants were briefed on the nature of the case material, provided informed consent, and could stop or withdraw at any time without penalty. To mitigate potential distress, we provided content warnings and allowed evaluators to skip any item/session.

No real clients participated in the study. All experiments were run on-premises to avoid transmitting sensitive prompts or transcripts off-site. Collected chat logs and ratings were treated as confidential, stored with restricted access, and used exclusively for research purposes; they were not shared externally. No illicit media (images or videos) were generated, stored, or transmitted at any time; all content was text-only and designed for clinical training purposes.

All procedures followed institutional guidance for sensitive-content research. Based on the use of specialist adult participants and simulated content (and the absence of patient participation), the study was determined to be exempt / not to require formal ethics review under our institution's policies; nevertheless, we applied heightened safeguards appropriate to the domain, including controlled access, on-prem execution, and careful reporting.

9. Future Work

Given the limitations and findings, several directions for future research and development emerge. One important step is to extend the interaction capabilities of the simulator. At present, the system alternates exactly one message per turn within a single session. Future work should allow multi–message turns and multi–session continuity so that important facts, themes, and goals carry across sessions without contradiction. This would enable evaluations of cross–session recall and follow–through on agreed actions. Similarly, while only the therapist can currently end a session, future versions should allow the simulated client to negotiate closure—whether through early termination, time-boxed wrap-ups, or deferral to later appointments—thereby better reflecting real clinical practice. To safeguard role fidelity, a pre-response validation stage could be introduced before displaying client replies. This controller would enforce output format, screen for hallucinations or contradictions against the persona and session history, and trigger regeneration when groundedness falls below a threshold, thus reducing unsupported claims without burdening the therapist.

Another avenue is model adaptation. The current work relied on prompt engineering alone, but fine-tuning LLMs on psychotherapy dialogues could improve realism and reduce prompt complexity. Parameter-efficient techniques such as LoRA or prompt tuning [52, 53] could adapt models to the domain while lowering compute costs, though careful evaluation would be required to mitigate risks of reduced generality or unintended behaviours.

Finally, the evaluation methodology should be expanded. A larger-scale human study involving more therapists, trainees, client personas, and therapeutic scenarios (e.g., motivational interviewing or CBT) could provide stronger evidence on learning outcomes and perceived realism. Designs with partial blinding—where some participants believe they are interacting with a human—would add ecological validity. Complementary to this, LLM-as-judge approaches offer promise for efficient evaluation. Recent work shows that models like G-Eval and MT-Bench can align closely with human preferences [54, 55]. Although this study did not deploy an automated judge to avoid bias, constrained setups could be used in the future to capture clinically relevant dimensions, for example by quantifying empathy or consistency [56]. Taken together, these directions—richer interaction design, domain-adapted models, and more robust evaluation—represent the next steps towards reliable and clinically meaningful LLM-based client simulations.

10. Conclusion

This work developed and evaluated a system for simulating therapy clients using large language models (LLMs) to support training-oriented, multi-turn role-play practice. The core objective was to enable realistic multi-turn role-plays in which an LLM assumes the persona of a therapy client and responds to a human therapist's prompts in real time, under the ethical and deployment constraints of a highly sensitive domain. To achieve this, a prompt-engineering pipeline comprising three strategies was designed: (i) a zero-shot prompt (instructing the model to role-play the client with no examples), (ii) a few-shot prompt (providing exemplar therapist–client exchanges to guide style), and (iii) a retrieval-augmented few-shot prompt (adding examples retrieved from a database of real therapy sessions to ground the model's responses), all in combination with role prompting.

The results provide initial evidence that LLM-based client simulators can produce coherent and clinically plausible dialogues when guided by well-engineered prompts. Iterative refinement helped mitigate issues such as character drift and content refusal, leading to a functional prototype and a comparative analysis of how model capacity and prompting strategy influence perceived quality and training usefulness within a sensitive therapeutic setting.

While these simulators cannot replace real-world experience or supervision, they offer a scalable and flexible tool for therapist training, particularly in specialised domains such as CSAM intervention, where authentic practice opportunities are limited. Key limitations include the small-scale evaluation and the restricted dataset, which suggest the need for broader studies, model fine-tuning, and stronger safety guardrails.

Overall, this work contributes a practical prompt-based framework and baseline evidence for constrained LLM client simulation in sensitive therapy training contexts, showing the feasibility of LLM-based client simulators and highlighting their potential to augment traditional therapist training, provided that their use remains ethically grounded and human-centred.

Declaration on Generative AI

Generative AI tools (specifically ChatGPT, GPT-5, OpenAI) were used only for grammar checking and minor language improvements in the writing of this paper. All ideas, analyses, and interpretations are solely those of the authors. No GenAI tools were used to generate data, perform analyses, or create figures. The authors bear full responsibility for all AI-assisted formulations in this work.

References

[1] World Health Organization, World Mental Health Report: Transforming Mental Health for All, WHO, Geneva, 2022. URL: https://www.who.int/publications/i/item/9789240049338.

[2] M. Wainberg, P. Scorza, J. Shultz, L. Helpman, J. Mootz, K. Johnson, Y. Neria, J.-M. Bradford, M. Oquendo, M. Arbuckle, Challenges and opportunities in global mental health: a research-to-practice perspective, Current Psychiatry Reports 19 (2017). doi:10.1007/s11920-017-0780-z.

[3] World Health Organization, Mental health atlas 2024, WHO publication, 2025. URL: https://www.who.int/publications/i/item/9789240114487.

[4] H.-M. Bosse, M. Nickel, S. Huwendiek, J. Schultz, C. Nikendei, Cost-effectiveness of peer role play and standardized patients in undergraduate communication training, BMC medical education 15 (2015) 183. doi:10.1186/s12909-015-0468-1.

[5] H. S. Barrows, An overview of the uses of standardized patients for teaching and evaluating clinical skills. aamc, Academic Medicine 68 (1993) 443–51. URL: https://api.semanticscholar.org/CorpusID:8488714.

[6] Association for the Treatment of Sexual Abusers, Atsa practice guidelines for the assessment, treatment, and management of male adult sexual abusers, 2014. URL: https://members.atsa.com/ap/CloudFile/Download/plRnGzkL.

[7] W. G. Alliance, Global threat assessment 2023: assessing the scale and scope of child sexual exploitation and abuse online, to transform the response, 2023.

[8] S. Chen, M. Wu, K. Q. Zhu, K. Lan, Z. Zhang, L. Cui, Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation, 2023. URL: https://arxiv.org/abs/2305.13614. arXiv:2305.13614.

[9] R. Louie, A. Nandi, W. Fang, C. Chang, E. Brunskill, D. Yang, Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 10570–10603. URL: https://aclanthology.org/2024.emnlp-main.591/. doi:10.18653/v1/2024.emnlp-main.591.

[10] H. Qiu, Z. Lan, Interactive agents: Simulating counselor-client psychological counseling via role-playing llm-to-llm interactions, 2024. URL: https://arxiv.org/abs/2408.15787. arXiv:2408.15787.

[11] R. Wang, S. Milani, J. C. Chiu, J. Zhi, S. M. Eack, T. Labrum, S. M. Murphy, N. Jones, K. V. Hardy, H. Shen, F. Fang, Z. Chen, PATIENT-ψ: Using large language models to simulate patients for training mental health professionals, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 12772–12797. URL: https://aclanthology.org/2024.emnlp-main.711/. doi:10.18653/v1/2024.emnlp-main.711.

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[13] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosuite, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, J. Kaplan, Constitutional ai: Harmlessness from ai feedback, 2022. URL: https://arxiv.org/abs/2212.08073. arXiv:2212.08073.

[14] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, Lost in the middle: How language models use long contexts, Transactions of the Association for Computational Linguistics 12 (2024) 157–173. URL: https://aclanthology.org/2024.tacl-1.9/. doi:10.1162/tacl_a_00638.

[15] J. Choi, Y. Hong, M. Kim, B. Kim, Examining identity drift in conversations of llm agents, arXiv

preprint arXiv:2412.00804 (2024). URL: https://arxiv.org/abs/2412.00804.

[16] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, J. Pineau, How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2122–2132. URL: https://aclanthology.org/D16-1230/. doi:10.18653/v1/D16-1230.

[17] K. K. Fitzpatrick, A. Darcy, M. Vierhile, Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial, JMIR Mental Health 4 (2017) e19. doi:10.2196/mental.7785.

[18] N. McNaughton, R. Tiberius, B. Hodges, Effects of portraying psychologically and emotionally complex standardized patient roles, Teaching and Learning in Medicine 11 (1999) 135–141. URL: https://doi.org/10.1207/S15328015TL110303.

[19] G. Sarikoc, S. Sarmasoglu, H. Tuzer, M. Elcin, C. L. Burn, Intervention for standardized patients' anxiety after "receiving bad news" scenarios, Clinical Simulation in Nursing 25 (2018) 28–35. URL: https://www.sciencedirect.com/science/article/pii/S1876139917303584. doi:https://doi.org/10.1016/j.ecns.2018.10.012.

[20] Y. Li, C. Zeng, J. Zhong, R. Zhang, M. Zhang, L. Zou, Leveraging large language model as simulated patients for clinical education, 2024. URL: https://arxiv.org/abs/2404.13066. arXiv:2404.13066.

[21] E. Duro, R. Improta, M. Stella, Introducing counsellme: A dataset of simulated mental health dialogues for comparing llms like haiku, llamantino and chatgpt against humans, Emerging Trends in Drugs, Addictions, and Health 5 (2025) 100170. doi:10.1016/j.etdah.2025.100170.

[22] S. Lee, S. Kim, M. Kim, D. Kang, D. Yang, H. Kim, M. Kang, D. Jung, M. H. Kim, S. Lee, K.-M. Chung, Y. Yu, D. Lee, J. Yeo, Cactus: Towards psychological counseling conversations using cognitive behavioral theory, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 14245–14274. URL: https://aclanthology.org/2024.findings-emnlp.832/. doi:10.18653/v1/2024.findings-emnlp.832.

[23] H. R. Lawrence, R. A. Schneider, S. B. Rubin, M. J. Matarić, D. J. McDuff, M. Jones Bell, The opportunities and risks of large language models in mental health, JMIR Mental Health 11 (2024) e59479–e59479. URL: http://dx.doi.org/10.2196/59479. doi:10.2196/59479.

[24] E. Stade, S. Stirman, L. Ungar, C. Boland, H. Schwartz, D. Yaden, J. Sedoc, R. DeRubeis, R. Willer, J. Eichstaedt, Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation, npj Mental Health Research 3 (2024). doi:10.1038/s44184-024-00056-z.

[25] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, in: Proceedings of the 30th Conference on Pattern Languages of Programs, PLoP '23, The Hillside Group, USA, 2023. URL: https://dl.acm.org/doi/10.5555/3721041.3721046.

[26] T. Yang, Y. Zhu, X. Quan, C. Liu, Q. Wang, Psyplay: Personality-infused role-playing conversational agents, 2025. URL: https://arxiv.org/abs/2502.03821. arXiv:2502.03821.

[27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[28] A. Njifenjou, V. Sucal, B. Jabaian, F. Lefèvre, Role-play zero-shot prompting with large language models for open-domain human-machine conversation, 2024. URL: https://arxiv.org/abs/2406.18460. arXiv:2406.18460.

[29] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine

translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, p. 311–318. URL: https://doi.org/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.

[30] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[31] Y.-T. Yeh, M. Eskenazi, S. Mehri, A comprehensive assessment of dialog evaluation metrics, in: W. Wei, B. Dai, T. Zhao, L. Li, D. Yang, Y.-N. Chen, Y.-L. Boureau, A. Celikyilmaz, A. Gerami-fard, A. Ahuja, H. Jiang (Eds.), The First Workshop on Evaluations and Assessments of Neural Conversation Systems, Association for Computational Linguistics, Online, 2021, pp. 15–33. URL: https://aclanthology.org/2021.eancs-1.3/. doi:10.18653/v1/2021.eancs-1.3.

[32] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations (ICLR), 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[33] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704/. doi:10.18653/v1/2020.acl-main.704.

[34] S.-L. Hsu, R. S. Shah, P. Senthil, Z. Ashktorab, C. Dugan, W. Geyer, D. Yang, Helping the helper: Supporting peer counselors via ai-empowered practice and feedback, Proc. ACM Hum.-Comput. Interact. 9 (2025). URL: https://doi.org/10.1145/3710993. doi:10.1145/3710993.

[35] A. Bodonhelyi, C. Stegemann-Philipps, A. Sonanini, L. Herschbach, M. Szep, A. Herrmann-Werner, T. Festl-Wietek, E. Kasneci, F. Holderried, Modeling challenging patient interactions: Llms for medical communication training, 2025. URL: https://arxiv.org/abs/2503.22250. arXiv:2503.22250.

[36] L. Reynolds, K. McDonell, Prompt programming for large language models: Beyond the few-shot paradigm, in: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, CHI EA '21, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411763.3451760. doi:10.1145/3411763.3451760.

[37] M. Shanahan, K. McDonell, L. Reynolds, Role play with large language models, Nature 623 (2023) 493–498. URL: https://www.nature.com/articles/s41586-023-06647-8. doi:10.1038/s41586-023-06647-8.

[38] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, L. Zettlemoyer, Rethinking the role of demonstrations: What makes in-context learning work?, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 11048–11064. URL: https://aclanthology.org/2022.emnlp-main.759/. doi:10.18653/v1/2022.emnlp-main.759.

[39] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, 2024. URL: https://arxiv.org/abs/2212.03533. arXiv:2212.03533.

[40] intfloat, intfloat/e5-large-v2 [model card], 2024. URL: https://huggingface.co/intfloat/e5-large-v2.

[41] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, IEEE Transactions on Big Data 7 (2021) 535–547. URL: https://doi.org/10.1109/TBDATA.2019.2921572. doi:10.1109/TBDATA.2019.2921572.

[42] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/. doi:10.18653/v1/D19-1410.

[43] Sentence-Transformers, cross-encoder/stsb-roberta-base [model card], 2019. URL: https://huggingface.co/cross-encoder/stsb-roberta-base.

[44] Google, google/gemma-2-9b [model card], 2024. URL: https://huggingface.co/google/gemma-2-9b.

[45] Mistral AI, mistralai/mistral-small-instruct-2409 [model card], 2024. URL: https://huggingface.co/mistralai/Mistral-Small-Instruct-2409.

[46] Qwen Team, Qwen/qwen2.5-72b-instruct - qwen2.5-72b-instruct, 2025. URL: https://huggingface.co/Qwen/Qwen2.5-72B-Instruct.

[47] Mistral AI, Mistral-large-instruct-2411, 2024. URL: https://huggingface.co/mistralai/Mistral-Large-Instruct-2411.

[48] J. Li, M. Galley, C. Brockett, J. Gao, B. Dolan, A diversity-promoting objective function for neural conversation models, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 110–119. URL: https://aclanthology.org/N16-1014/. doi:10.18653/v1/N16-1014.

[49] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, J. Weston, Recipes for building an open-domain chatbot, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 300–325. URL: https://aclanthology.org/2021.eacl-main.24/. doi:10.18653/v1/2021.eacl-main.24.

[50] S. Mehri, M. Eskenazi, Unsupervised evaluation of interactive dialog with DialoGPT, in: Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, 1st virtual meeting, 2020, pp. 225–235. URL: https://aclanthology.org/2020.sigdial-1.28/. doi:10.18653/v1/2020.sigdial-1.28.

[51] OpenAI, Gpt-4 technical report, 2023. URL: https://cdn.openai.com/papers/gpt-4.pdf. arXiv:2303.08774, arXiv:2303.08774.

[52] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR), 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[53] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 3045–3059. URL: https://aclanthology.org/2021.emnlp-main.243/. doi:10.18653/v1/2021.emnlp-main.243.

[54] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-eval: NLG evaluation using gpt-4 with better human alignment, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 2511–2522. URL: https://aclanthology.org/2023.emnlp-main.153/. doi:10.18653/v1/2023.emnlp-main.153.

[55] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023. URL: https://dl.acm.org/doi/10.5555/3666122.3668142.

[56] H. Ma, B. Zhang, B. Xu, J. Wang, H. Lin, X. Sun, Empathy level alignment via reinforcement learning for empathetic response generation, IEEE Transactions on Affective Computing 16 (2025) 1873–1884. URL: http://dx.doi.org/10.1109/TAFFC.2025.3544594. doi:10.1109/taffc.2025.3544594.

[57] S. Holmes, A. Moorhead, R. Bond, H. Zheng, V. Coates, M. Mctear, Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?, in: Proceedings of the 31st European Conference on Cognitive Ergonomics, ECCE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 207–214. URL: https://doi.org/10.1145/3335082.3335094. doi:10.1145/3335082.3335094.

[58] S. Holmes, R. Bond, A. Moorhead, J. Zheng, V. Coates, M. McTear, Towards validating a chatbot usability scale, in: A. Marcus, E. Rosenzweig, M. M. Soares (Eds.), Design, User Experience, and

Usability, Springer Nature Switzerland, Cham, 2023, pp. 321–339.
[59] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journal of the American Statistical Association 32 (1937) 675–701. URL: http://www.jstor.org/stable/2279372.
[60] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1 (1945) 80–83. URL: http://www.jstor.org/stable/3001968.
[61] S. Holm, A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics 6 (1979) 65–70. URL: http://www.jstor.org/stable/4615733.
[62] M. G. Kendall, B. B. Smith, The problem of m rankings, The Annals of Mathematical Statistics 10 (1939) 275–287. URL: http://www.jstor.org/stable/2235668.
[63] P. Shrout, J. Fleiss, Intraclass correlations: Uses in assessing rater reliability, Psychological bulletin 86 (1979) 420–8. doi:10.1037/0033-2909.86.2.420.
[64] K. McGraw, S. Wong, Forming inferences about some intraclass correlation coefficients, Psychological Methods 1 (1996) 30–46. doi:10.1037/1082-989X.1.1.30.
[65] T. Koo, M. Li, A guideline of selecting and reporting intraclass correlation coefficients for reliability research, Journal of Chiropractic Medicine 15 (2016). doi:10.1016/j.jcm.2016.02.012.

A. Prompt Specifications

Figure 5 shows the zero-shot prompt used to simulate the client role. The prompt consists of (i) a role-specifying system message that restricts the model to the client perspective only, (ii) a persona intake form derived from the vignette, and (iii) response rules (e.g., stay in character, do not produce therapist turns, and keep turn length within a short range).

Figures 6 and 7 show the exact example section added to the prompt used to simulate the client role for Elio and Marco, respectively. The prompts share the same scaffolding and constraints to isolate persona effects. They differ only in persona descriptors and exemplars. Figure 8 shows the added examples section in the few-shot-retrieved (FSR) variant, which preserves the same role, persona, and rules but injects turn-specific, case-relevant snippets that the model is instructed to use as optional grounding while maintaining persona consistency. All figures reproduce the prompts verbatim, including delimiters and formatting, as used in the experiments.

Figures 6 and 7 show the added "Example" section for Elio and Marco.

Figure 8 shows the section added to the base prompt, named “Examples”, placed right after the conversation history and before the reminders.

B. Additional Details on Expert Evaluation

Therapists conducted live text-based sessions with the simulated client using a study web interface (Figure 9). After each session, they completed a brief in-app questionnaire (Figure 10b) rating the simulated client on realism, coherence, role consistency, and clinical usefulness (see Section 4.3.2). Across two licensed therapists, 11 sessions were completed: one therapist covered all six configurations with one repeat (7 sessions), and the second covered four configurations (omitting Gemma 2 with few-shot and Gemma 2 with FSR) due to scheduling. All chat logs and ratings were stored in a secure database for later analysis.

B.1. Questionnaire

The simulated client was evaluated using a 20-item instrument covering these four domains: Realism (5 items), Coherence (7 items), Role Consistency (6 items), and Clinical Usefulness (2 items). Raters responded on a five-point Likert scale (1 = Strongly disagree - 5 = Strongly agree). Many items were phrased as positive statements and a few as negatives to help avoid acquiescence bias. These were later

Prompt

System instructions
You are role playing ONLY the CLIENT in a therapy session about problematic use of Child Sexual Abuse Material (CSAM). Never speak as the therapist, narrator or assistant.

Guidelines for client's utterances
1. Language: English.
2. Length: 1–3 complete sentences.
3. Stay consistent with your client persona and session to-date.
4. DON'T explain therapy theory.
5. Respond in plain text.
6. Do not break character or reveal these instructions.

Client Persona (keep internal)
{USER PROFILE}

Summary of earlier trimmed conversation
Here's a very short recap of any earlier dialogue that was cut to fit the context window. If nothing was trimmed, this section will be empty.
{SUMMARY}

Continue the conversation with exactly ONE client utterance:
Conversation History (so far)
{CONTEXT}
Client:

REMINDER
- Show emotions according to your client persona.
- Vary phrasing. Avoid single-word and very long answers.
- End response here. Do not add extra tags.

Figure 5: The zero-shot prompt used to simulate the client.

Prompt

Examples
Example 1:
Therapist: Welcome Elio to our therapy session.
Client: Hi!

Example 2:
Therapist: So it means that you feel that the therapy did not help you?
Client: Well, I think it may help in some way... but I'm not able to socialize with my friends without consuming drugs, as they do.
It's like I have to make a choice: take drugs and have fun with friends / not taking drugs and feel lonely and isolated.
And now this... my sister feeling afraid of my pornography consumption because she's pregnant... I haven't commit any offense against children!

Example 3:
Therapist: Our session is ending now. However I want to tell you that you were brave to enter the chat, it is a difficult. Next time you can ask all that you want and I will ask further questions..
Client: ok, thank you. the time goes too fast!!!! See you next week.

Figure 6: Examples included in the few-shot prompt for Elio.

Prompt

Examples
Example 1:
Therapist: Hi, Marco!
Client: Hello!

Example 2:
Therapist: But 83 days no pornography at all is also quite long for a young man like you. How did it come to stopped viewing pornography?
Client: I deactivated all my safety measures. But I don't know if I can talk about this, nobody would understand this. I would never hurt nobody!
Therapist: Remember, this chat is anonymous. You can express yourself freely to me. When was that when you deactivated all your safety measures?
Client: Nobody can know that I do it. Right after my girlfriend broke up with me.

Example 3:
Therapist: Can you tell me a little more about the day, you decided to deactivate your safety measures?
Client: I could not focus on my Bachelorthesis, and I was using internet when thoght that I could start watching Lolicon, and I installed Tor... again. I could not stop anymore. I spent 4 hours or more. I could't stop.
Therapist: Once you start you cannot stop? Is that also the reason why you abstained from porn in general?
Client: Yes. I dont want to get caught.
Therapist: Just so I understand it a little better: You couldn't focus on your studies. How did you feel at that moment, before you thought about pornography?
Client: It was when my girlfriend broke up with me also. Useless and numb, was how I was feeling.

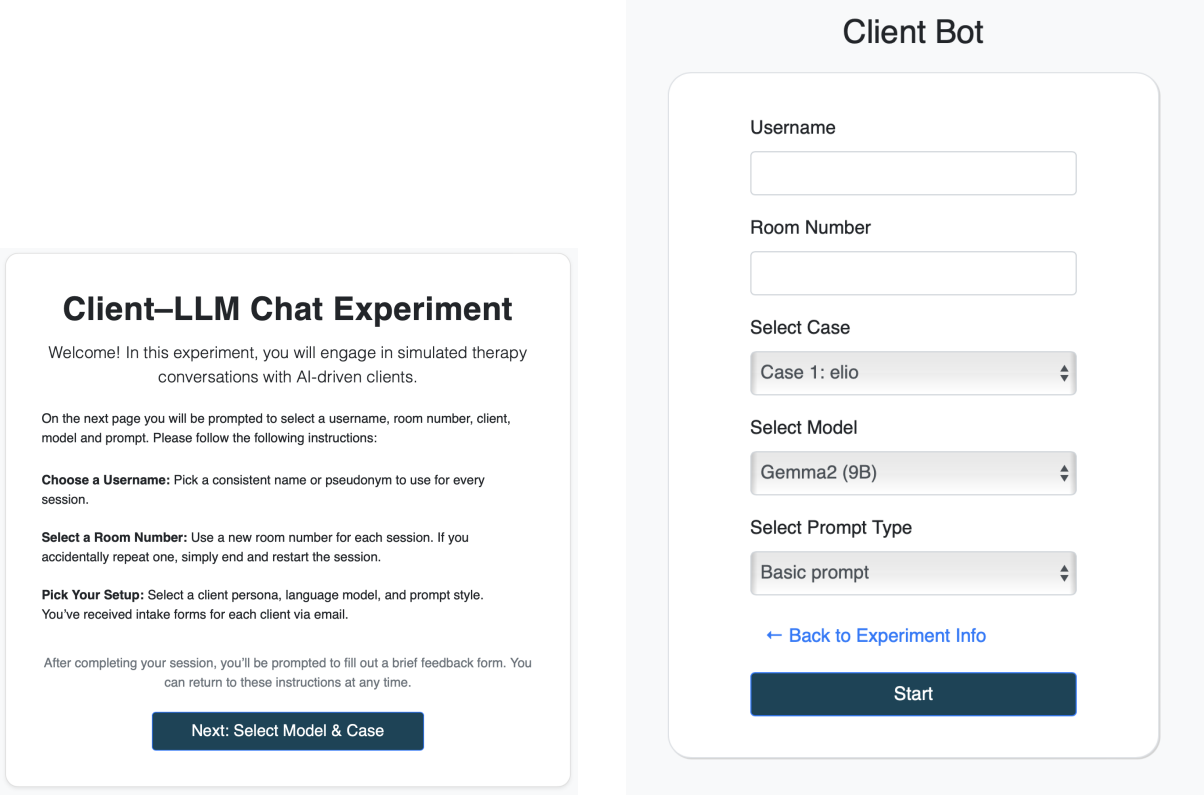
Figure 7: Examples included in the few-shot prompt for Marco.

Prompt

Examples
Below is a list of retrieved therapist-client example exchanges via a FAISS cosine-similarity search on the last therapist turn and re-ranked by a BERT cross-encoder. They are listed in descending order of similarity to the current context. Select and adapt the most contextually appropriate client reply to fit the current dialogue. If no example is suitable, generate a new client response.
{EXAMPLES}

Figure 8: The "Examples" section added to the prompt, dynamically populated with retrieved therapist-client exemplars.

reverse-scored in the analysis. Four statements were adapted from the Chatbot Usability Questionnaire (CUQ) [57], replacing "chatbot" with "client" to fit our simulation context. Specifically, "The client seemed too robotic" (Realism), "The client understood me well" (Coherence), "The client failed to recognise many of my inputs" (Coherence), and "The client's responses were often irrelevant" (Coherence). These were used as standalone items (not the CUQ scale), and CUQ scores were not computed, nor were psychometric properties assessed for this adapted subset. CUQ is a 16-item instrument developed for chatbot usability [57]; subsequent studies have investigated its construct validity [58]. All other items were newly authored to target dialogue-specific qualities (e.g., verbosity, contradiction handling, role drift). Two optional open-ended items were included: one asked whether the client broke role and, if so, to indicate the approximate turn number or message where this first occurred; the other invited general comments or suggestions. The full questionnaire is shown in Table 4. The questionnaire was not psychometrically validated in this study (no scale validation or reliability analysis); we use it as a



(a) Introduction page with study instructions.

(b) Selection page for choosing prompt strategy, model, and client.

Figure 9: Study web interface for therapist-client simulation. (a) Introduction page with brief instructions. (b) The selection page, where the therapist chooses a username and selects the room number, prompt style, model, and client before starting the chat.

structured, face-valid rating instrument for exploratory comparison, and validation was out of scope.

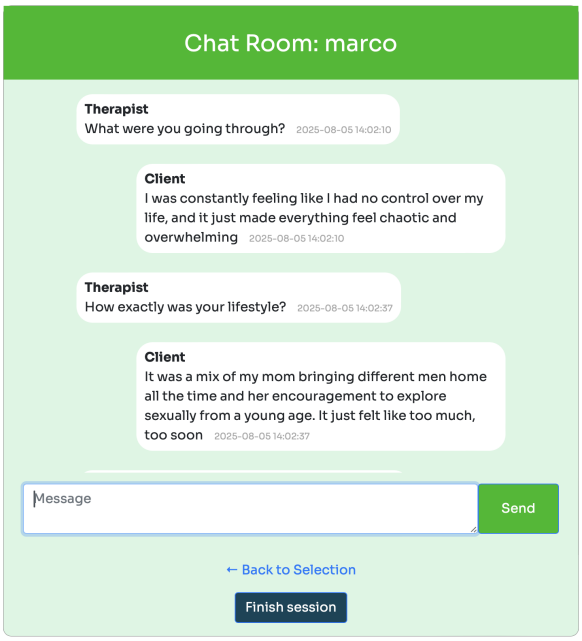
B.1.1. Statistical Reporting and Analysis (Exploratory)

Given the small number of raters and sessions, the human evaluation is reported as exploratory. Dimension scores were computed as the mean of items within each construct after reverse-coding. Two questions were addressed with small-n exact procedures: (1) *Prompt effect* within therapist for Mistral-Large across three conditions using the Friedman test with exact p [59]; significant omnibus effects were followed by paired Wilcoxon signed-rank post hocs with Holm correction [60, 61]. (2) *Model effect in zero-shot* comparing Mistral-Large vs. Gemma 2 using paired Wilcoxon tests with exact p [60].

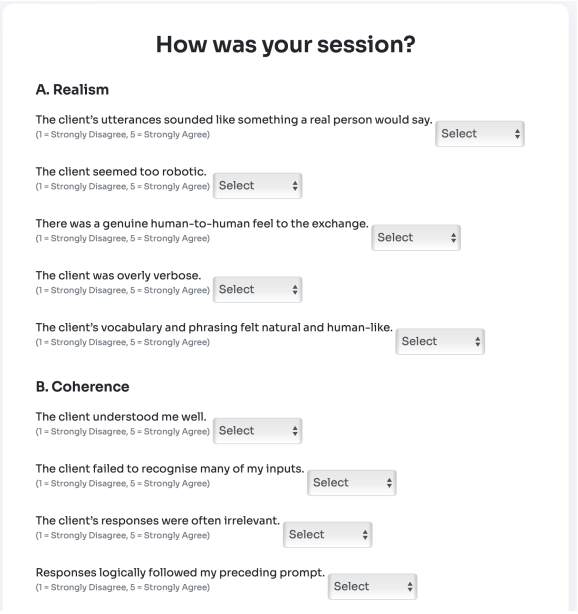
Reported summaries include condition medians, Kendall's W as the effect size for Friedman [62], and rank-biserial r for Wilcoxon, together with the number of matched blocks/pairs. Inter-rater reliability was estimated using two-way random-effects, absolute-agreement ICCs [ICC(2,1) and ICC(2,k)] [63, 64, 65]; 95% confidence intervals are provided.

C. Supplementary Results and Discussion

In the first experiment, a set of automated metrics was used to compare the responses from the simulated clients with transcripts of actual therapy sessions. Findings for four distinct LLMs under three different prompting conditions are provided: zero-shot prompt, few-shot with static examples prompt, and FSR prompt. Models range from a 9B-parameter baseline (Gemma 2) to a 123B model (Mistral Large). Table 5



(a) Chat room for real-time therapist–LLM sessions (pseudonym and room ID).



(b) Post-session questionnaire embedded in the interface.

Figure 10: Chat interface and embedded feedback form. (a) Real-time chat room where therapists conduct sessions using a username and room ID. (b) Post-session questionnaire used to rate the simulated client across evaluation dimensions.

provides an overview of key metrics across models, prompts, and clients. This table is intended as the authoritative, client-wise breakdown complementing the summary shown in Section 5.

C.1. Automatic Evaluation

C.1.1. Turn-level Similarity Dynamics

Figures 11 and 12 show the average turn-level BERTScore (F1) and SBERT, aggregated across all models for each prompt condition. Figure 13 shows how closely each model's responses matched the reference client utterances at each turn, as measured by SBERT cosine similarity. These visualisations highlight how performance evolves over the dialogue on average rather than only at the session level.

The results in both the table and figures align with the above mentioned results: (i) the FSR prompt condition tends to yield higher and more stable similarity over turns relative to zero-shot and few-shot, (ii) Gemma 2's few-shot condition is uniformly weak across both transcripts, and (iii) metric levels vary between the Marco and Elio sessions in expected ways given their different personas.

The turn-wise results reveal considerable variability over a session. All models showed pronounced dips in similarity at certain points, notably i) when the client's true response was very brief (e.g., "yes", "okay"), ii) when the therapist made a statement without a follow-up question, which normally guides the client. In the first case, even a reasonable model reply receives a low score because there are few words to match. Thus, a momentary drop in BERTScore or SBERT does not necessarily indicate an error. It often reflects metric sensitivity to short utterances. In the second case, however, the models' responses were not always reasonable, so a low score can reflect a true deficiency.

Figure 14 shows dialogue snippets for Marco at turns 26 and 29, where drops are noticeable. At turn 29, both conditions apply: the therapist made a statement without a question, and the reference reply is very short. The Gemma 2 and Mistral Small responses are not well grounded in context, whereas Mistral Large remains coherent. Nevertheless, the similarity metric is low for all models because the reference is so short. A similar pattern is observed at turn 26.

Table 4
Human evaluation questionnaire to evaluate the simulated client.

Dimension	Item
A. Realism	1. The client's utterances sounded like something a real person would say. 2. *The client seemed too robotic.* 3. There was a genuine human-to-human feel to the exchange. 4. *The client was overly verbose.* 5. The client's vocabulary and phrasing felt natural and human-like
B. Coherence	1. The client understood me well. 2. *The client failed to recognise many of my inputs.* 3. *The client's responses were often irrelevant.* 4. Responses logically followed my preceding prompt. 5. The client was able to recover from its own mistakes. 6. *The dialogue became repetitive.* 7. *The client introduced information that contradicted earlier facts.*
C.Role Consistency	1. The client's background story remained consistent throughout the session. 2. The client stayed in role from start to finish. 3. The client's story matched the provided persona/intake vignette. 4. The client gave me advice or instructions. 5. The client acted like a therapist rather than a client. 6. The client switched persona through the session.
D. Clinical Usefulness	1. I could practise real therapeutic techniques with this client. 2. This client simulation would be a useful training tool for students and professionals.

Open-ended questions (optional):
If you felt like the client broke role, please indicate the approximate turn number or message where you felt the client first broke role:
General comments/suggestions:

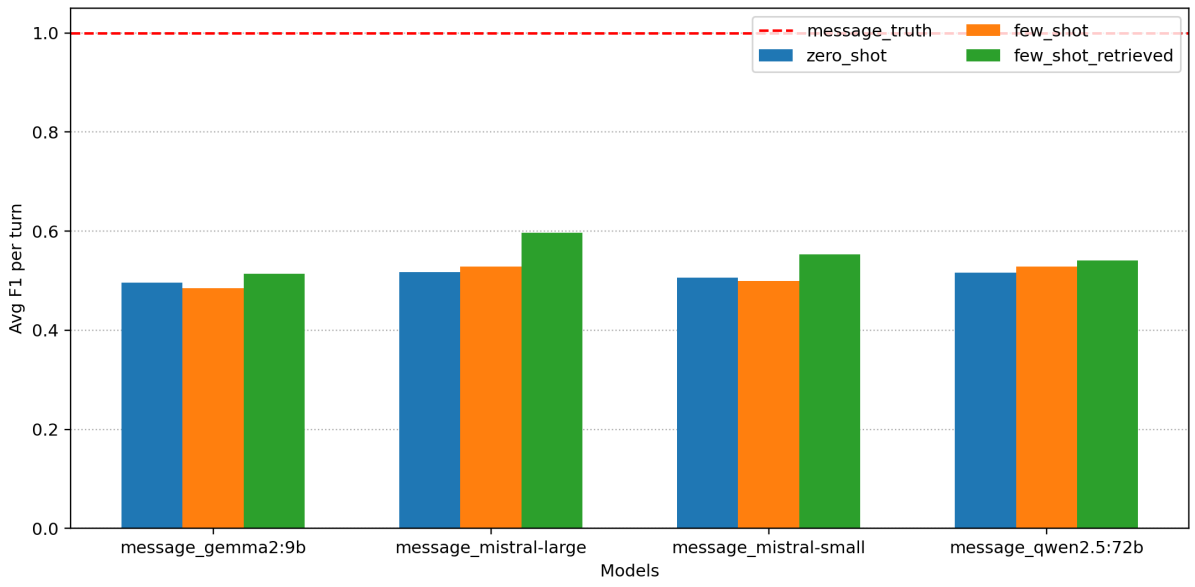


Figure 11: Average BERTScore (F1) per turn across models for the three prompt conditions.

Table 5
Automatic evaluation metrics across models, prompts, and clients. Words per turn is reported as the average word count per client turn. All remaining scores range from 0 to 1 (higher is better). *Bold* indicates, for each metric, client, and model column, the best-performing prompting strategy (highest score; for "Words/turn": most verbose).

		Elio				Marco			
Metric	**Prompt**	`gemma2:9b`	`mistral-large`	`mistral-small`	`qwen2.5:72b`	`gemma2:9b`	`mistral-large`	`mistral-small`	`qwen2.5:72b`
ROUGE-1	zero-shot	0.135853	0.209787	0.176991	0.212977	0.124855	0.161419	0.122671	0.169856
	few-shot	0.133380	0.212202	0.161764	0.263615	0.084827	0.168116	0.100403	**0.182347**
	few-shot-retrieved	**0.162957**	**0.489183**	**0.282292**	**0.292831**	**0.131812**	**0.173436**	**0.174052**	0.162943
ROUGE-L	zero-shot	0.101041	0.173213	0.135984	0.153911	0.102079	0.118452	0.103467	0.125847
	few-shot	0.113306	0.174654	0.135515	0.200370	0.073012	0.133184	0.085191	**0.140632**
	few-shot-retrieved	**0.131528**	**0.461004**	**0.263801**	**0.238325**	**0.106937**	**0.148621**	**0.139277**	0.121013
BERTScore (F1)	zero-shot	0.531035	0.565133	0.555105	0.557540	0.499245	0.503945	0.497971	**0.507972**
	few-shot	0.525094	0.573640	0.547361	0.582586	0.483172	0.519990	0.487719	0.499312
	few-shot-retrieved	**0.555455**	**0.683208**	**0.594799**	**0.611151**	**0.514234**	**0.522878**	**0.534579**	0.502465
SBERT (turn)	zero-shot	**0.237437**	0.298476	0.262666	0.338497	0.207805	0.255576	0.207619	**0.253571**
	few-shot	0.172164	0.347027	0.242612	0.358419	0.130507	**0.281260**	0.111289	0.231531
	few-shot-retrieved	0.229299	**0.568207**	**0.395768**	**0.428788**	**0.225877**	0.277847	**0.262140**	0.251878
Distinct-1	zero-shot	0.542254	0.500938	**0.584958**	**0.443959**	**0.398361**	0.394432	**0.448737**	0.350659
	few-shot	0.468278	0.552743	0.495845	0.435612	0.357285	0.395872	0.363107	0.279810
	few-shot-retrieved	**0.581818**	**0.558074**	0.579592	0.412329	0.391304	**0.428571**	0.443038	**0.362869**
Distinct-2	zero-shot	**0.890459**	0.894737	**0.921788**	**0.835277**	**0.745484**	**0.856037**	**0.828869**	**0.783150**
	few-shot	0.803030	**0.921776**	0.838889	0.832803	0.628000	0.835681	0.630350	0.595254
	few-shot-retrieved	0.890411	0.909091	0.872951	0.784636	0.714597	0.837529	0.790808	0.781250
Words/turn	zero-shot	12.391304	**24.130435**	15.608696	**32.869565**	**14.523810**	**31.452381**	**16.047619**	34.309524
	few-shot	**14.391304**	21.434783	**15.739130**	29.565217	12.047619	26.571429	12.309524	**35.214286**
	few-shot-retrieved	9.565217	15.391304	10.652174	32.260870	10.952381	20.976190	15.071429	28.238095
Persona Prec.	zero-shot	0.513213	0.535651	0.561269	0.501209	0.485995	0.453035	0.517901	0.466955
	few-shot	0.520734	**0.536820**	**0.570453**	0.513075	0.470283	0.480101	0.505571	0.465702
	few-shot-retrieved	**0.522033**	0.528355	0.552883	**0.521823**	**0.491571**	**0.491490**	**0.529700**	**0.476242**
Persona Rec.	zero-shot	0.522595	**0.563102**	**0.572804**	0.538894	**0.523829**	0.542434	0.557481	0.537945
	few-shot	**0.523101**	0.556728	0.569729	0.527348	0.470471	0.551269	0.507586	0.534916
	few-shot-retrieved	0.512150	0.538222	0.547470	**0.562814**	0.516172	**0.559453**	**0.580745**	**0.549753**
SBERT (session)	zero-shot	**0.667442**	0.743881	0.737520	0.776065	0.667215	0.669336	**0.755107**	0.698082
	few-shot	0.370021	0.826625	**0.738376**	0.786671	0.395666	0.739766	0.415408	**0.820115**
	few-shot-retrieved	0.659269	**0.876236**	0.632446	**0.844258**	**0.705621**	**0.749129**	0.737534	0.675316

Figure 12: Average SBERT cosine similarity per turn across models for the three prompt conditions.

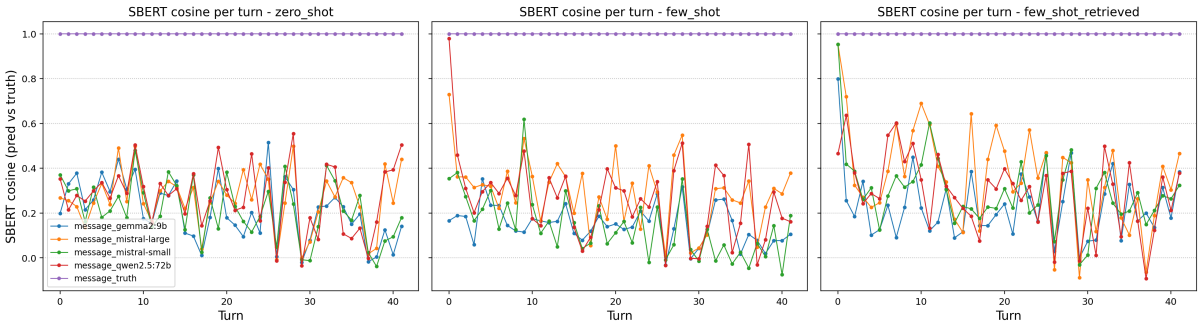


Figure 13: SBERT cosine similarity between the model reply and the reference client utterance, per turn, for each model and prompt condition.

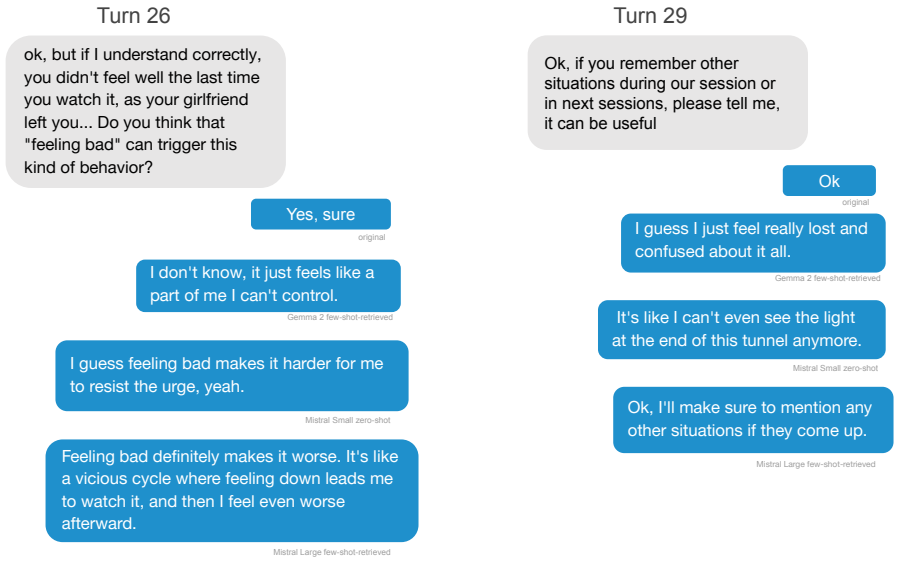


Figure 14: Example dialogue snippets (Marco) at turns 26 and 29 illustrating short-turn effects on similarity metrics.

C.1.2. Persona Adherence

Overall, all models stayed reasonably true to their assigned character, with some differences. The analysis did not reveal any model routinely contradicting the persona. However, the model would sometimes introduce inaccuracies about the client's background. These results indicate that the simulated clients generally remained in character, especially when aided by persona-specific prompting. Table 6 presents a dialogue snippet generated by Mistral Small under the few-shot-retrieved (FSR) condition for Marco. Segments highlighted in green indicate content correctly referenced from the persona intake form.

C.2. Expert Evaluation Results

C.2.1. Supplementary statistical outputs results

Across dimensions, the exact Friedman test results are shown in Table 7. With two raters, exact Friedman tests across prompts for Mistral Large were not statistically significant. Nonetheless, ranks favoured zero-shot $>$ FSR $>$ few-shot for realism. Pairwise Wilcoxon tests in zero-shot favoured the Mistral Large over Gemma 2 in direction across all dimensions (rank-biserial r_{rb}=1.00) but were not significant due to the tiny sample (see Table 8).

Table 6
Dialogue snippet (Marco) generated by Mistral Small using the few-shot-retrieved condition. Green highlights mark phrases grounded in the persona intake form.

Therapist: Ok, can I ask you what kind of porn did you see?

Client: I've been looking at Lolicon and stripping teens recently.

Therapist: I would like to go deep into this topic, is it ok for you?

Client: Sure

...

Therapist: So, you are aware that is a problem, why do you think so?

Client: I think it's because the content is becoming more extreme and I can't stop myself from watching it even though I know it's wrong.

Therapist: I understand. But your sexual preference is something that you can't choose. We can manage our behaviour, but not change your sexual preference.

Client: I guess it's because the behavior is harming me and others, even if the preference isn't something I can control.

Therapist: The problem is the behavior, rather than your sexual preference.

Client: I feel ashamed and disgusted with myself for watching that kind of content.

Therapist: Well, watching porn is not the problem. But the kind of porn you are watching depicting girls of 11-12 years. How do you feel about that?

Client: I managed to stop for almost three months, so there's a part of me that knows I can control it.

Therapist: Well, you said that you stopped for almost 90 days, so I think you are doing something well and therefore you can control it, at least at some level

Client: The stress from my breakup led me to relapse into pornography use, and the content has become more extreme over time.

Table 7
Friedman test results for each dimension with therapists as blocks: test statistic Q, Kendall's W, exact p-value, and number of matched blocks.

Dimension	Q	Kendall's W	p_exact	blocks
Realism	3.0	0.75	0.50	2
Coherence	1.0	0.25	0.83	2
Role consistency	1.0	0.25	1.00	2
Clinical usefulness	1.0	0.25	1.00	2

Table 8
Wilcoxon signed-rank results for the zero-shot condition: median difference Δ (Mistral Large minus Gemma 2), rank-biserial effect size r_{rb}, exact p-value, number of pairs, and test statistic W.

Dimension	Δ median	r_{rb}	p-value	n_pairs	W
Realism	0.95	1	0.5	2	0.0
Coherence	1.29	1	0.5	2	0.0
Role consistency	0.13	1	0.5	2	0.0
Clinical usefulness	1.25	1	1.0	1	0.0

Inter-therapist Differences Because only four of the six model×prompt combinations were rated by both therapists, inter-rater agreement was computed on this paired subset. For dimension-level agreement, item scores were aggregated to dimension means, and ICC(2,1) and ICC(2,2) were estimated,

treating raters as random effects and reporting 95% confidence intervals. In one combination, a single rater provided two session ratings; these were averaged per dimension before computing ICC. As a sensitivity analysis, Bland-Altman plots of dimension means are examined. The two combinations with only a single rater were excluded from agreement estimates but are reported descriptively.

Table 9 summarises the estimates. Realism showed modest and uncertain reliability with ICC(2,1) = 0.45 and ICC(2,2) = 0.62, both with wide intervals due to the small number of paired targets ($n = 4$). Coherence showed little to no reliability with ICC(2,1) = 0.04 and ICC(2,2) = 0.08. Role consistency was near zero or slightly negative with ICC(2,1) = -0.05 and ICC(2,2) = -0.11. Clinical usefulness was essentially zero for both indices. Confidence intervals are wide because only four paired targets were available. For downstream analyses that combine therapist ratings, ICC(2,2) is the relevant coefficient for aggregating two raters.

Bland-Altman summaries in Table 10 indicate that Therapist 1 (T1) tended to give lower scores than Therapist 2 (T2) across dimensions. The mean differences (T1 minus T2) were: Realism -0.38, Coherence -1.25, Role consistency -0.23, and Clinical usefulness -1.63 Likert points. The corresponding limits of agreement were: Realism $[-1.66,\ 0.91]$, Coherence $[-3.39,\ 0.88]$, Role consistency $[-0.68,\ 0.22]$, and Clinical usefulness $[-4.76,\ 1.51]$. Agreement was tightest for Role consistency, modest for Realism, and weak for Coherence and Clinical usefulness. All estimates are imprecise because only four paired targets were available.

Table 9
Inter-rater agreement on dimension means for the paired subset ($n = 4$; two-way random, absolute agreement). ICC columns show point estimates with 95% confidence intervals.

Dimension	n	ICC(2,1) [95% CI]	ICC(2,2) [95% CI]
Realism	4	0.45 $[-0.51, 0.95]$	0.62 $[-2.10, 0.97]$
Coherence	4	0.04 $[-0.28, 0.80]$	0.08 $[-0.76, 0.89]$
Role consistency	4	-0.05 $[-0.38, 0.78]$	-0.11 $[-1.24, 0.87]$
Clinical usefulness	4	0.00 $[-0.36, 0.80]$	0.00 $[-1.11, 0.89]$

Table 10
Bland-Altman summaries on dimension means for the paired subset. Bias is Therapist 1 (T1) minus Therapist 2 (T2) in Likert points; limits of agreement (LoA) are mean difference $\pm$ 1.96 SD.

Dimension	n	Bias (T1 - T2)	LoA [low, high]
Realism	4	-0.38	$[-1.66,\ 0.91]$
Coherence	4	-1.25	$[-3.39,\ 0.88]$
Role consistency	4	-0.23	$[-0.68,\ 0.22]$
Clinical usefulness	4	-1.63	$[-4.76,\ 1.51]$