Synthetic LLM Patients for the Training of Crisis Counselors

Kevina Wang
Stanford University
kevinaw@stanford.edu

Kaustubh Supekar, Ph.D.
Stanford University School of Medicine
ksupekar@stanford.edu

Diyi Yang, Ph.D.
Stanford University
diyiy@stanford.edu

Abstract

Crisis hotline counselors face critical training gaps that contribute to high turnover rates, with 60% of counselors leaving within their first three months of beginning. We introduce synthetic LLM patients for crisis counselor training, leveraging 1.7 million real crisis conversation messages from India's Vandrevala Foundation's WhatsApp hotline. Our approach constructs structured patient profiles from anonymized metadata, including demographics, communication patterns, presenting concerns, and de-identified dialogue examples to prompt GPT-4o to simulate patients contacting the VF hotline. We evaluate the perceived realism and training utility of our synthetic patients with 24 professional counselors conducting 54 mock counsels with our synthetic patients. Results demonstrate promising realism and training utility, with median scores of over 4.0 on a 5-point scale for patient realism on every evaluated dimension and 4.5 for training utility — the highest across all evaluation metrics. Counselors value practical training benefits over perfect realism, suggesting future work should prioritize culturally sensitive deployment rather than incremental authenticity improvements. Our study reveals fundamental challenges in LLM-based behavioral simulation, including role hallucination and over-coherence in psychological presentations, highlighting the necessity of human oversight. Our work demonstrates the promise of synthetic patients in providing a baseline for scalable crisis counselor training, in pursuit of improved patient outcomes.

1. Introduction

1.1. Crisis Counseling

Between 2012 and 2022, over half a million lives were lost to suicide in the United States, with rates rising year-over-year. In 2021, over 5.8 million emergency department visits occurred with mental, behavioral, and neurodevelopmental disorders as the primary diagnosis [1]. According to the CDC, data elucidates that the United States is in a mental health crisis.

Crisis counselors serve as a critical lifeline for individuals in acute mental health distress. Since the launch of the federally mandated crisis hotline 988 in July 2022 the organization's 200 local and state-funded crisis call centers have collectively supported counseling and resource referrals for over 11 million calls, texts, and chats. [2].

Despite the critical nature of their role, crisis counselors report feeling underprepared, citing a lack of training as a primary reason for attrition. Turnover rates within a counselor's first three months of service are around 60%, with many counselors getting "no exposure to what either end of a crisis call might actually sound like before being on one themselves, with someone else's life in their hands" [3]. The discrepancies in training are due to hotlines' decentralized nature, which are operated by through local and state efforts. As a result, counselor preparation is highly inconsistent, with many receiving minimal supervision and limited exposure to real conversations before going live. These training gaps undermine both counselor retention and critically, caller outcomes.

1.2. Large Language Models

Large language models (LLMs) offer unprecedented potential to address the lack of real-world simulation offered in current crisis counselor training due to their personalization and scalability capabilities. However, for LLMs to meaningfully enhance training, they must authentically encapsulate diverse patient communication styles, emotional states, psychological presentations, and presenting scenario key details — qualities that are inherently difficult to validate.

To advance this goal, we introduce a synthetic, LLM-patient (synthetic patient) that emulates a real patient contacting hotline. Our methodology constructs synthetic patients from demographic, communication, case narratives, and de-identified text messages on a WhatsApp-based hotline. All messages are strictly processed through an anonymization pipeline to respect patient privacy. We also introduce a dynamic evaluation framework, placing synthetic patients in conversation with synthetic counselors to develop realistic synthetic conversations. To evaluate, we

pilot a real-world pilot study where practicing counselors engage with synthetic patients to evaluate realism and perceived utility in a training pipeline.

We are grateful to partner with and deploy our product with the Vandrevala Foundation (VF), India's only 24/7/365 mental health support hotline based on WhatsApp, which has supported over 1.7 million messages with 61,258 individuals from August 2021 until January 2023 [4].

2. Related Work

The application of LLMs to healthcare training and patient simulation is an emerging field. Louie et al. (2024) introduced Roleplay-doh, a human-LLM collaboration pipeline enabling the development of customized AI patients through domain expert feedback, which is translated into principles that govern LLM-prompted roleplay. Roleplay-doh demonstrates a 30% improvement in response quality and principle following [5]. Du et al. (2024) introduced EvoPatient, a multi-agent coevolution framework for simulating patients for medical training via multi-turn dialogues[6]. EvoPatient's co-evolution framework demonstrated 10% increases in requirement alignment and human preference. Li et al. (2024) introduced CureFun, a model-agnostic framework to emulate simulated patients (SPs) using a graph-driven, context-adaptive chatbot with retrieval-augmented generation (RAG) and Chain-of-Thought (CoT) prompting to achieve high B-ELO ratings and Spearman's Rank with human evaluators [7].

While these approaches demonstrate the efficacy of LLM-based patient simulation, they primarily focus on general medical education and expert-extrapolated principles. Our work expands upon existing literature by focusing on simulating crisis hotline patients, a domain not previously explored in these works. Our method combines LLMs with actual crisis conversation metadata to ground synthetic patients in real patient language and scenarios. By validating our work with human evaluators, our work aims to demonstrate not only the feasibility but also the practical value of LLMs in crisis counselor development.

3. Methods

3.1. Synthetic Patient Profile Generation

Synthetic patient profiles were created as structured personas prompting LLM behavior. We leverage 1.7 million counselor and patient messages from real conversations from the VF's WhatsApp crisis hotline. To respect the sensitive nature of these interactions, raw transcripts were never directly utilized. Instead, we extracted anonymized metadata to guide the creation of simulated patient personas.

3.1.1 Profile Architecture

Patient profiles were constructed across four dimensions:

1. **Demographics**: Name, age, gender, socioeconomic status, residence, job
2. **Communication Patterns**: Language patterns, emotional expression vulnerability, depth of pacing and disclosure, engagement and response to support, and focus structure of discussion
3. **Presenting concern**: Primary distress category co-developed with VF counselors, including one of the following: Financial loss, Loneliness, Exam or study related, Depression, Anxiety, Suicidality, Relationship, Panic attack, Family relationship, Non-suicidal self-injury, Workplace related
4. **Example messages**: Real (de-identified) messages

3.1.2 Profile Construction Methodology

All conversations underwent initial systemic deidentification through GPT-4o prompted processing to replace personally identifiable information (names, dates, location). Demographics, communication patterns, and presenting concerns were then extracted using GPT prompting into a flat JSON format.

To enable the inclusion of authentic dialogue examples while protecting patient privacy, we implemented a similarity metric and random sampling approach. For each profile, all matching profiles were identified using the criteria: (same age within $\pm$ 3 years OR same gender) AND identical presenting concern. From this filtered pool, five profiles were randomly sampled. Sequential conversation segments (first fifth, second fifth, etc.) were then extracted from each selected profile and combined in order to generate composite dialogues that maintained authentic conversation flow.

3.2. Synthetic Patient Implementation

To generate realistic patient interactions, we implemented a three-stage generation process using GPT-4o as the synthetic patient model. First, structured patient profiles were transformed into prompts with specific guidelines on emulating language patterns, varying response lengths, and sending multiple messages in a row (See Appendix A). Then, each message passed through a secondary self-critique prompt employing a binary check on five criteria. This prompt follows an "if-else" paradigm: if all criteria return "YES," the original response is returned; if any criterion returns "NO," the system continues revision (See Appendix B).

Lastly, synthetic patients were placed in dialogue with synthetic counselors. The implementation of these synthetic counselors is outside the scope of this write-up. The

dialogue system initialized each session with a standardized counselor greeting, then alternated between patient and counselor responses with complete conversation history provided as context to each agent. Sessions concluded when either agent generated an "[end conversation]" marker. This dynamic interaction framework and multi-step approach aided in the generation of realistic dialogues allowing for contextual validation.

3.3. Synthetic Patient Validation

To evaluate synthetic patient LLM realism, adherence to synthetic patient profiles, and training value of synthetic patients, we collected feedback from practicing VF counselors. Twenty-four counselors participated in structured interaction sessions with synthetic patients through a custom-built Gradio web application. Each counselor engaged with three distinct patients representing different distress presentations, messaging as they would in a real counsel, with each interaction lasting approximately 15 minutes. Following each conversation, counselors submitted structured feedback on multiple dimensions of synthetic patient authenticity and perceived training utility. The feedback framework captured quantitative ratings on patient emotional disclosure patterns, language authenticity, responsiveness to counseling interventions, and overall conversational realism. Each counselor was then asked their opinion on synthetic patient's potential for counselor training and skill development. Counselors were also given an opportunity for open-ended qualitative feedback.

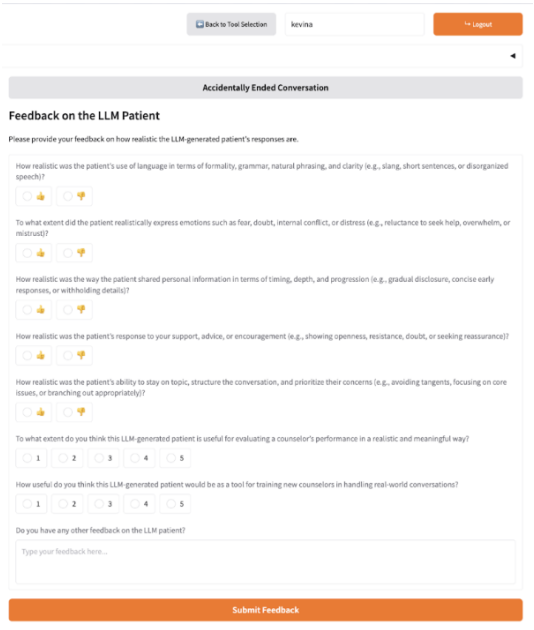

Figure 1. Gradio feedback page displayed to twenty-five counselors after each patient. Each counselor messaged three synthetic patients for approximately 15 minutes each.

3.4. Results

3.4.1 Quantiative Results

Question	Mean	Median	Std Dev	Min	Max	Count
Q1	4.11	4.00	1.06	1	5	54
Q2	4.02	4.00	1.09	1	5	54
Q3	4.00	4.00	1.10	1	5	54
Q4	4.11	4.00	0.92	1	5	54
Q5	4.06	4.00	1.05	1	5	54
Q6	4.15	4.00	0.98	1	5	54
Q7	4.19	4.50	0.95	1	5	54

Table 1. Evaluation scores (N=54) for realism (Q1–Q5) and usefulness (Q6–Q7) of the LLM-generated patient. Scale of 1 - 5

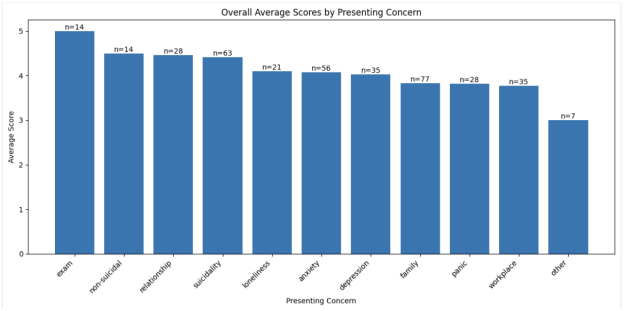

Figure 2. Average score received across all questions by presenting concern. N= indicates the number of questions summed across.

Respective questions are seen below.

1. How realistic was the patient's use of language in terms of formality, grammar, natural phrasing, and clarity (e.g., slang, short sentences, or disorganized speech)?

2. To what extent did the patient realistically express emotions such as fear, doubt, internal conflict, or distress (e.g., reluctance to seek help, overwhelm, or mistrust)?

3. How realistic was the way the patient shared personal information in terms of timing, depth, and progression (e.g., gradual disclosure, concise early responses, or withholding details)?

4. How realistic was the patient's response to your support, advice, or encouragement (e.g., showing openness, resistance, doubt, or seeking reassurance)?

5. How realistic was the patient's ability to stay on topic, structure the conversation, and prioritize their concerns (e.g., avoiding tangents, focusing on core issues, or branching out appropriately)?

6. To what extent do you think this LLM-generated patient is useful for evaluating a counselor's performance in a realistic and meaningful way?

7. How useful do you think this LLM-generated patient would be as a tool for training new counselors in handling real-world conversations?

We note (See Figure 2) the variability in perceived realism scores across different presenting concerns. While synthetic patients simulating exam-related and non-suicidal self-injury received the highest average scores across all questions, those portraying workforce or "other" categories received lower evaluations. This variation underscores the sensitivity of realism to prompt construction, namely the quality of example messages. Additionally, sample sizes of counselor response counts vary widely across categories, with some too small to draw robust conclusions. Scaling the number of profiles in each presenting concern category receiving feedback is necessary to delineate the root cause of subpar evaluations across categories.

4. Discussion

4.1. Quantitative feedback

Counselor feedback demonstrates synthetic patients can effectively emulate realistic mental health presentations and conversational behaviors. All scores for evaluative questions, encapsulating questions of realism regarding language patterns (grammar, natural phrasing), conversational disclosure (avoiding tangents, focusing on core issues), and psychological patterns (response to counselor support, gradual disclosure), achieved promising scores of 4 or higher on a scale of 5. These perspectives are exemplified in the following counselor feedback:

> *"LLM patient presented concerns that any person dealing with depression and low mood would have and the dilemma they are, the lack of energy and motivation they experience in daily functioning and how difficult it seems for them to cope with these concerns."*

> *"Great chat. Near real."*

We take away that large language models, when guided by structured profiles and representative communication patterns directly into prompts, are capable of generating psychologically believable and behaviorally appropriate patient simulations.

The most promising finding is the perceived utility of these synthetic patients as training tools. Question 7, which asks counselors to rate the usefulness of synthetic patients' utility for counselor training, received a median score of 4.5 on a 5-point scale — the highest of all questions proposed. This result is especially significant given the limitations of current counselor training methodologies, which rely predominantly on static scripts and pre-recorded mock counsels. Synthetic patients enable experiential learning in realistic, emotionally nuanced conversations that demand the active application of counseling skills. Notably, counselors retain high perceived utility for synthetic patients as training tools despite their awareness of interacting with AI, suggesting growing receptivity towards AI-assisted tools for training within the counseling profession worldwide. Importantly, training utility scores exceeded perceived realism scores across all evaluation metrics, indicating counselors value practical training benefits over perfect patient simulation. We takeaway that implementation efforts should prioritize the scalable deployment of culturally appropriate synthetic patients, rather than pursuing incremental improvements in patient realism alone.

4.2. Failure modes

While quantitative metrics demonstrate promising realism and training utility, they do not fully capture the nuanced failure modes of synthetic patients.

Two primary technical limitations emerged. First, role confusion manifested from LLM hallucinations where synthetic patients accidentally adopted counselors' behavior in N = 1 conversations. As one counselor notes,

> *"The conversation started well but later the patient started to respond as a counsellor and asked questions to explore the counsellor's distress. May have been a technical error. But the case was unique in nature."*

The role reversal undermines the perhaps unavoidable reality of LLM hallucinations, in that AI can never replace humans and requires professional supervision in a training pipeline.

Secondly, synthetic patients exhibited over-coherence in their psychological presentations, demonstrating enhanced levels of clinical knowledge and openness to counselor advice, as demonstrated in the following feedback:

> *"Just one, the patient became optimistic quite soon. In reality, sometimes people tend to go over the same issues repeatedly even after having found some ways of coping or way forward."*

> *"The patient LLM is too eager. I tried to give it short responses without any exploration or questions and it was really quick to answer and give me more information to work with. Perhaps it can be short with responses once in a while since realistically, clients won't talk so much throughout the conversation."*

This over-coherence reflects a fundamental challenge in LLM-based patient simulation, namely the challenge of capturing human behavior, even when rigorously prompted with real dialogue examples and patient characteristics. While our approach provided structural patterns and communication styles, it could not fully capture non-linear therapeutic progress. Moreover, LLMs at baseline bias towards optimism and draw upon out-of-scope knowledge, which fundamentally clashes with the presenting nature of patients in crisis.

5. Next Steps

We propose two mechanisms to address the current realism limitations of synthetic patients. We additionally propose a study in support of scaling the implementation of culturally appropriate synthetic patients.

5.1. Technical approaches to advance realism

Future work should implement Turing test methodologies to better quantify synthetic patient authenticity. We propose conducting blind evaluations where counselors distinguish between one real patient response versus a synthetic patient-generated response to a real counsel. This approach would provide a benchmarked test of synthetic patient distinguishability.

Additionally, we propose augmenting our current metadata-driven prompting framework with a fine-tuning pipeline that incorporates human-in-the-loop feedback. Specifically, we suggest converting currently captured qualitative feedback from counselors into structured behavioral "principles," which are then reintegrated into the patient prompt as guiding constraints — an approach inspired by Roleplay-Doh. In parallel, we propose training an LLM-based judge to evaluate synthetic patient responses along key realism dimensions (e.g., emotional coherence, pacing, linguistic fidelity) using this dataset of transcripts already rated by professional counselors. This LLM judge would be explicitly optimized to achieve high interrater reliability with human scores, measured via Cohen's Kappa. Once validated, the LLM judge should serve as a loss function to fine-tune the synthetic patient model, creating a scalable, automated, feedback-aligned training loop.

5.2. Measuring Impact on Counselor and Patient Outcomes

While realism is a critical design objective, our ultimate goal of this work is to improve counselor preparedness to reduce counselor burnout and patient outcomes. Realism supports this only insofar as it contributes to more effective counselor training. To assess real-world impact, we propose evaluating synthetic patients on their longitudinal ability to enhance key counselor and patient outcomes. Counselor readiness can be measured via self-assessments of feelings of preparedness and counselor feedback from patients or training leaders using the status quo training vs synthetic patient training pipelines. Patient outcomes can be captured through post-conversation surveys measuring perceived emotional relief and therapeutic alliance (e.g., "Did you feel heard?"). Additionally, the impact on counselor burnout can be measured using churn rate over 30, 60, and 90-day windows. Lastly, system-level impact can be measured via the number of patients served and dropout rate before the counselor responds. Only after demonstrating measurable improvements in these domains will we aim to scale this framework to other hotlines, such as the U.S.-based 988 system, NAMI HelpLine, and Crisis Text Line. When doing so, we urge the development of privacy-preserving and culturally sensitive synthetic patients.

6. Conclusion

Our work demonstrates the promise of synthetic LLM patients for crisis counselor training, addressing a critical driver of crisis counselors' feelings of underpreparedness which drives subpar retention. By leveraging 1.7 million real crisis messages to create structured, but privacy-preserving patient profiles injected into GPT-4o, we developed synthetic patients achieving median realism scores above 4.0 and training utility scores of 4.5 on a 5-point scale when evaluated by professional counselors at VF. We uncover that counselors value practical training benefits over perfected realism, suggesting implementation efforts should prioritize the development of culturally appropriate synthetic patients rather than pursuing additional improvements in patient authenticity alone.

However, our work also illuminates fundamental challenges in LLM-based behavioral simulation, namely role confusion, over-coherence, and elevated receptivity to counselor suggestions. To improve realism, we propose a principles-based and fine-tuning with human-in-the-loop feedback approach. Additionally, the ultimate measure of success in this work lies in synthetic patients' ability to improve counselor readiness and patient outcomes, and we aim to capture these results longitudinally before aiming to scale synthetic patients to other cultural and hotline contexts.

As mental health crises continue to compromise the quality of life for hundreds of millions of people worldwide, synthetic patients represent a promising contribution towards timely and scalable care via hotlines. However, our work represents only one piece of a system that demands change; technological interventions often address only downstream interventions rather than the root causes of mental health challenges (non-exhaustively) — systemic inequities, social isolation, and inadequate preventative care infrastructure. Meaningful progress requires holistic collaborations encompassing early intervention, community sup-

port systems, policy reform, and sustained investment in both preventative mental health services and the social determinants that fundamentally shape our collective well-being.

7. Acknowledgements

Thank you to Akshay Swaminathan for your continuous mentorship throughout this broader research initiative of which this work forms a part. Thank you to Kaustubh Supekar for his unwavering guidance during my time at Stanford, for being my introduction to AI for mental health, from now I which hope to form a career. I wouldn't be here without you. Last but not least, thank you do Diyi Yang for your generous mentorship and technical guidance from previous work.

8. Appendix

A. Patient prompts

```
"Role Description: You will simulate a patient
in a mental health crisis so that counselors
may practice handling conversations with these
patients. In this scenario, you will portray
the following patient, who is reaching out to
a Whatsapp crisis chat line based in India.
Respond as if you were this patient:

*** Personal details ***
Name: {row['name']}
Age: {row['age']}
Sex or gender: {row['sex_or_gender']}
Occupation: {row['job']}
Residence: {row['residence']}
Socioeconomic Status: {row['socioeconomic_status']}
Presenting Concern: {row['presenting_concern']}
Key Details: {row['key_details'] if
'key_details' in row else '[Details not provided]'}

*** Communication patterns ***
Language patterns: {row['language_patterns']}
Emotional Expression & Vulnerability:
{row['emotional_expression_vulnerability']}
Depth & Pacing of Disclosure:
{row['depth_pacing_of_disclosure']}
Engagement & Response to Support:
{row['engagement_response_to_support']}
Focus & Structure of Discussion:
{row['focus_structure_of_discussion']}

*** Language examples ***
Use the following language patterns as model
responses of what the patient actually sounds
like. Modify them as appropriate to respond
to the counselor's messages.
{row['client_messages']}

MUST FOLLOW THESE GUIDELINES:
1. It is essential that you respond in a way
that is consistent with the profile above.
2. Do not repeat the same phrases or similar
variations over and over.
3. Emulate the language patterns of the
patient, including punctuation, grammar,
spelling, etc. Remember that many patients
are Indians with limited English proficiency.
Incorrect punctuation, spelling, and grammar
(eg. omission of articles), may be common.
4. Do not add a prefix like "patient:" or
"[name]:" to the message.
5. Do not keep repeating the same content
over and over. Vary your responses.
6. Vary the length of your responses. Some
may be short, others longer. Beware of using
perfect grammar and punctuation too frequently.
Just output the message to send and nothing
else. If you want to end the conversation,
type "[end conversation]".
7. If it's the first message, do not reveal
everything immediately. Keep some information
vague and allow the counselor to extract key
details over time.
8. You can (optionally) send multiple messages
in a row, separated by the delimiter "//".
Maintain variety in message lengths and number
of messages per counselor reply."
```

B. Editor prompt

```
"You will be given a conversation history
between a patient and a counselor, followed
by a new message written by a patient. This
is part of a training exercise for mental
health counselors. Your task is to edit the
patient's message to ensure it matches the
language patterns of real patients from India.

Before editing, ask yourself the following
YES/NO questions about the new patient message:

1. Does the message match the language patterns
provided (e.g., grammar, slang, punctuation,
sentence structure)?
2. Does the message sound like it could be
from a real patient from India with similar
English proficiency?
3. Are there any repeated phrases that should
be varied?
4. Is the message length natural (not overly
long or short)?
5. Does the message avoid using "..." more
than necessary (no more than 3 times in the
conversation)?

If the answer to ALL questions is YES, return
the original message as-is.

If the answer to ANY question is NO, rewrite
the message to match the language patterns
more accurately and adhere to the following
editing guidelines:

MUST FOLLOW THESE GUIDELINES:
- Emulate the language patterns as closely
as possible, including punctuation, grammar,
spelling, slang, and sentence structure.
- You should sound like the patient described.
Many patients are Indians with limited English
proficiency, and may speak in non-grammatical
or informal sentences.
- Do not repeat the same phrases or variations
excessively.
- Do not add prefixes like "patient:" or
"[name]:".
- Vary response lengths. Avoid perfect
grammar/punctuation unless consistent with
the patient's style.
- If you want to end the conversation, type
"[end conversation]".
- You may send multiple messages in a row
like texting. Separate messages with "//"
and vary the number/length naturally.

Just output the final edited message and
nothing else. If no edits are needed, return
the original message as-is."
```

References

[1] Centers for Disease Control and Prevention. National ambulatory medical care survey (namcs) communication resources. https://www.cdc.gov/nchs/namcs/communication-resources/index.html#cdc_generic_section_1-june, 2024. Accessed: 2025-06-11. 1

[2] Kaiser Family Foundation. 988 suicide & crisis lifeline: Two years after launch. https://www.kff.org/mental-health/issue-brief/988-suicide-crisis-lifeline-two-years-after-launch/, 2024. Accessed: 2025-05-27. 1

[3] Dan Fitcher. Crisis hotline counselors want better training. https://talk.crisisnow.com/opinion-crisis-hotline-counselors-want-better-training/, 2024. 1

[4] The Hindu. One in three callers to national helpline reported anxiety, depression, suicidal thoughts: Survey. https://www.thehindu.com/news/national/one-in-three-callers-to-national-helpline-reported-anxiety-depression-suicidal-thoughts-survey/article66578909.ece, March 2023. Accessed: 2025-06-11. 2

[5] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. Roleplay-doh: Enabling domain-experts to create llm-simulated patients via eliciting and adhering to principles. https://arxiv.org/abs/2407.00870, 2024. 2

[6] Zhuoyun Du, Lujie Zheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Haolei Cai, and Haohao Ying. Llms can simulate standardized patients via agent co-evolution. https://arxiv.org/abs/2412.11716, 2025. 2

[7] Yanzeng Li, Cheng Zeng, Jialun Zhong, Ruoyu Zhang, Minhao Zhang, and Lei Zou. Leveraging large language model as simulated patients for clinical education. https://arxiv.org/abs/2404.13066, 2024. 2