

PatientHub: A Unified Framework for Patient Simulation

Sahand Sabour TszYam NG Minlie Huang

The CoAI Group, DCST, Institute for Artificial Intelligence,
Tsinghua University, Beijing, China

sahandfer@gmail.com, aihuang@tsinghua.edu.cn

Abstract

As Large Language Models increasingly power role-playing applications, simulating patients has become a valuable tool for training counselors and scaling therapeutic assessment. However, prior work remains fragmented: existing approaches rely on incompatible, non-standardized profiles, prompts, and evaluation metrics, hindering reproducibility, fair comparison, and reuse. We introduce PatientHub, a unified and modular framework that standardizes the creation, simulation, and evaluation of LLM-based patients. Our framework provides 16 patient simulators, a graph-based orchestrator for multi-turn, multi-session interactions, and a configurable LLM-as-a-judge evaluator that supports multiple rubric types. Via our command-line interface, users can generate patient profiles, run simulations, and apply rubric-driven evaluation at the turn and session level. To demonstrate PatientHub’s utility, we compare several supported simulators under a shared interaction protocol and showcase its extensibility by prototyping a new simulator variant with minimal additional code. By consolidating existing work into a single reproducible pipeline, our framework eliminates much of the infrastructure overhead that currently fragments this research and accelerates the development of new methods. Our code and data are publicly available via <https://github.com/Sahandfer/PatientHub>.

1 Introduction

Mental health disorders such as depression and anxiety are prevalent and create a sustained need for timely, accessible, and appropriate support. Yet access to such support remains limited due to severe workforce constraints: the global median is only 13 mental health workers per 100,000 people, with rates considerably lower in low-income countries (World Health Organization, 2025). Meeting this need at scale requires progress along two complementary directions: training more counselors and

peer supporters (Steenstra et al., 2025; Wang et al., 2024b; Louie et al., 2025), and developing LLM-based supporters that provide scalable, low-cost support under appropriate safeguards (Sabour et al., 2023; Chen et al., 2023; Zhou et al., 2025). However, both directions face practical bottlenecks.

First, many training programs still rely heavily on static materials (e.g., textbooks) that cannot provide interactive practice or feedback (Lin et al., 2026). A viable alternative to this is standardized patients (i.e., trained actors), in which students practice necessary skills (e.g., active listening) in low-risk settings; this approach can be effective but is costly and difficult to scale (Yao et al., 2020). Moreover, it is highly challenging to diversify the patient population across different demographic and cultural backgrounds and symptom profiles.

Second, data scarcity and safety risks constrain the development of LLM-based supporters (Yang et al., 2025a; Archiwaranguprok et al., 2025). Creating publicly available, diverse, high-quality therapy conversations for training LLMs is difficult because of privacy concerns, ethical constraints, and the potential cost of expert involvement. In addition, rigorously assessing LLM-based supporters requires a safe simulated environment, as involving real human users carries safety risks (e.g., failures can lead to harmful advice), which makes large-scale evaluation challenging.

A growing line of work is leveraging LLM-based simulated patients for training, evaluation, and data generation (Lee et al., 2024, 2025a,b). Compared with standardized patients, these simulators offer lower marginal cost, greater personalization, and rapid scaling across diverse cases. In addition, they have shown strong potential for generating large-scale, high-quality synthetic datasets for training and evaluating LLM-based supporters (Yang et al., 2025a; Zhao et al., 2024; Liu et al., 2025a).

However, existing work remains scattered across papers and codebases, with substantial variation

in how simulated patients are specified (e.g., profile structures), how conversations are orchestrated (single- vs. multi-turn; single- vs. multi-session), and how responses/outputs are evaluated (e.g., automatic evaluation, expert judgments, or task-specific rubrics). In principle, this diversity is a strength, as it enables researchers to explore various applications of simulated patients. Yet, this fragmentation makes it difficult to reproduce prior work, compare results across different methods, and reuse existing components (e.g., profiles, prompts, evaluation rubrics). These capabilities are essential for building reliable benchmarks, identifying which design choices matter, and facilitating the development of novel methods by reducing duplicated effort.

To this end, we present PatientHub, a unified, modular framework that standardizes the creation, simulation, and evaluation of simulated patients. PatientHub provides a common interface for implementing diverse mental health agents, a graph-based orchestrator that composes multi-turn interactions as configurable events, and a rubric-driven LLM-as-a-judge evaluator for downstream analysis. Through our command-line interface, researchers can generate patient profiles, run simulations across heterogeneous methods, and apply standardized evaluation, while extending the framework with their own profiles, prompts, and metrics. Our contributions are summarized as follows:

- Unify 16 patient simulators, covering a broad range of existing approaches, under a shared abstraction of agents, events, generators, and evaluators, enabling fair cross-method and cross-model comparison that was previously hindered by incompatible codebases.
- Provide a configurable LLM-as-a-judge evaluator supporting four rubric paradigms at both the turn and session levels.
- Support rapid development of new methods, which we demonstrate by prototyping a PATIENT- Ψ variant with minimal code.
- Release the framework (MIT license) with documentation and an installable package, lowering the barrier to entry.

2 Related Work

2.1 Patient Simulation Methods

LLM-based patient simulation has emerged as a flexible alternative to traditional rule-based sim-

ulators, leveraging modern LLMs’ role-playing capabilities. Prior work varies primarily in how patient behavior is specified (e.g., profile design), what interaction protocol is assumed (e.g., single- vs. multi-session), and how desired attributes (e.g., clinical faithfulness) are enforced. Existing methods are primarily prompt-based, conditioning generation on structured profiles and explicit constraints (Wang et al., 2024b; Louie et al., 2024; Wang et al., 2024a), and may incorporate explicit latent states or turn-level updates to reduce drift over long conversations (Yang et al., 2025b; Wang et al., 2025a). Fewer approaches directly train a patient simulator; to the best of our knowledge, Eeyore (Liu et al., 2025b) remains the only method to date to fine-tune a depression simulator via expert-in-the-loop supervision and preference optimization. Task settings also impact design, as simulators are often designed to meet specific clinical needs. Simulators for screening and psychiatric assessment ground responses in symptom checklists or interview schemas (Wang et al., 2025b; Lee et al., 2025a; Liao et al., 2024), whereas counseling-oriented simulators emphasize human-like variability (Du et al., 2025; Zhang et al., 2026). Despite this diversity, these methods are built in isolation across incompatible codebases, which PatientHub consolidates under a shared abstraction.

2.2 Applications of Patient Simulation

Patient simulators provide repeatable, low-risk role-play settings for training and assessment. In clinical education, they offer a scalable alternative to standardized patients, enabling trainees to practice essential skills such as empathy and active listening (Steenstra et al., 2025; Wang et al., 2024b) and providing contextualized feedback (Louie et al., 2025; Campillos-Llanos et al., 2021; Holderried et al., 2024). Moreover, they serve as standardized testbeds for benchmarking supporter agents while avoiding the risks and cost of real-user evaluation (Sabour et al., 2025; Bi et al., 2025; Chiu et al., 2024). Simulated interactions also enable synthetic datasets for training and analysis (Yang et al., 2025a; Lee et al., 2024; Mandal et al., 2026). Across these applications, rubric-driven *LLM-as-a-judge* evaluation has become the practical mechanism for measuring efficacy, realism, and safety (Liao et al., 2024; Emerson et al., 2025; Yosef et al., 2024; Zhu et al., 2025). Yet, rubrics and protocols differ across studies, motivating the standardized evaluator that PatientHub provides.

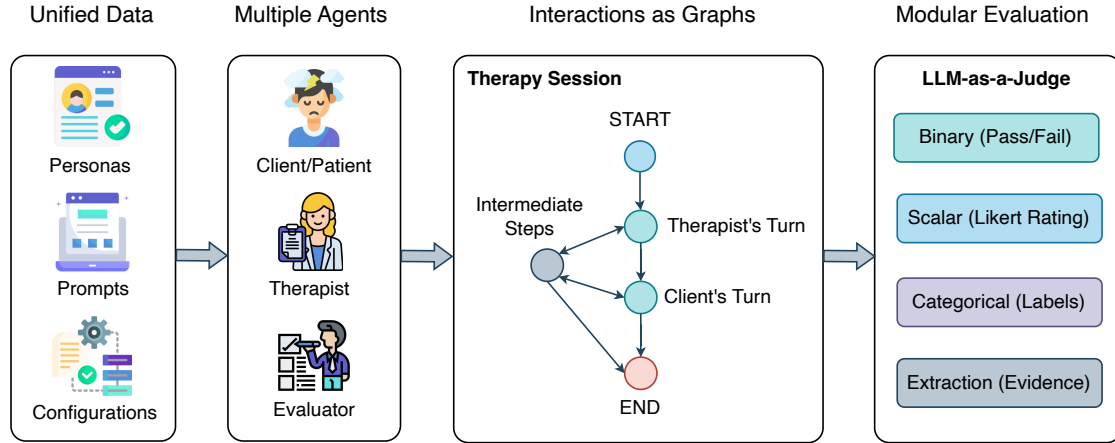


Figure 1: An Overview of PatientHub. Our framework includes a unified collection of personas, prompts, and configurations from existing work; support for several agent roles for interactions; orchestration of dialogue flows/events as directed graphs; and cross-method benchmarking through a modular LLM-as-a-Judge Evaluator supporting four fundamental evaluation paradigms.

3 PatientHub

PatientHub is a modular framework designed for the standardized creation, simulation, and evaluation of patient-centered interactions. Rather than specifying a single fixed design, we define a small set of composable abstractions (including agents, events, generators, and evaluators), which enable researchers to generate comparable simulations across diverse healthcare settings, from psychiatric assessment to CBT counseling. Concretely, users can specify (i) who participates in an interaction, (ii) how the interaction unfolds, and (iii) how its inputs and outputs are evaluated, then run the entire pipeline from a single command-line interface.

3.1 Core Components

Clients. Client agents simulate patients in healthcare or counseling interactions, spanning both diagnosed patients and individuals seeking support. Each client is specified by (i) a persona profile, (ii) a response-generation specification constraining output, and optionally (iii) an internal latent state that evolves over the dialogue to model emotional shifts or symptom progression. PatientHub currently supports 16 simulators from existing work (Table 3); we reused the original prompts and profiles verbatim from the published papers or codebases, and only re-implemented the surrounding orchestration code to fit our abstractions.

Therapists. Therapist agents serve as counterparts to clients and use the same high-level inter-

face. Alongside a prompt-driven therapist with optional chain-of-thought and a rule-based ELIZA (Weizenbaum, 1966), our framework also includes methods such as CAMI (Yang et al., 2025a) for motivational interviewing. Researchers can vary the therapist as they do the client, using the same profiles, events, and evaluators.

Events. Events orchestrate agent interactions, represented as graphs that specify protocols for turn-taking, termination criteria, and session structure. This design enables researchers to define workflows aligned with their experimental objectives. For instance, as shown in Figure 1, the *Therapy Session* event runs as a finite-state workflow, including initialization, therapist turn, client turn, and an end/reminder branch; the event is terminated upon reaching the turn limit or an explicit end signal, and all turns are logged to a unified JSON schema for downstream analysis.

Generators. Generators synthesize the client profiles that drive simulations, constructing them from seed sources such as conversation excerpts, clinical questionnaires, or structured disorder descriptions (Wang et al., 2024a, 2025a; Lai et al., 2026). By controlling the distribution of generated profiles, researchers can reproducibly scale simulations across diverse demographics and symptom presentations.

Evaluators. PatientHub provides a standardized *LLM-as-a-judge* abstraction that covers two judges: *profile judge* for generated profiles and *conversation judge* for full conversations, applied at the

session or turn level. Either judge can run online during a simulation or offline on saved logs. Each evaluation dimension is defined in a prompt template under one of four paradigms: (i) Binary (deterministic pass/fail; e.g., “Did the patient mention self-harm?”); (ii) Scalar (Likert ratings; e.g., “Rate the consistency from 1 to 5.”); (iii) Categorical (labels; e.g., “How realistic are the patient’s responses? good/neutral/bad”); and (iv) Extraction (evidence spans; e.g., “Highlight statements that contradict the given profile.”). This abstraction facilitates instruction customization while keeping the interface uniform across tasks and methods.

3.2 Usage

PatientHub’s workflow is accessible through a single command-line interface. Users can generate profiles from seed records, simulate an interaction between agents, evaluate sessions or profiles with a judge, and create a new agent implementation that becomes immediately available to the framework. Moreover, as patient simulators typically define profiles in incompatible formats, reusing a persona across methods otherwise requires manual reconstruction. Our framework includes an `adapt` command that maps a persona to another method’s schema. By automating this conversion, PatientHub enables a single persona to drive heterogeneous simulators and supports controlled cross-method comparison. Hydra can override all settings, and an optional web interface supports interactive, human-in-the-loop sessions.

3.3 Technical Details

PatientHub is implemented in Python with an emphasis on reproducibility and modularity. We use Hydra to manage configurations across datasets, agents, and experimental settings, Burr to represent and execute events as graphs, and LiteLLM as a unified interface for LLM calls. Structured outputs are validated with Pydantic schemas, and prompts are authored as Jinja-rendered YAML templates over JSON character records. New components are added by inheriting a base class (e.g., `BaseClient`) and registering it in the framework, at which point it becomes immediately available to the CLI.

4 Evaluation

We demonstrate PatientHub’s utility through a single end-to-end workflow over a shared set of personas: we i) create client profiles through generation and adaptation, validating them with the profile

judge; (ii) simulate and compare several methods under a common protocol, scored by the conversation judge; and (iii) extend the framework with a new method. Each stage demonstrates a core capability; we aim to show that PatientHub standardizes and executes the full pipeline, not to benchmark simulators or identify a single best method.

4.1 Experimental Setup

Therapists. We adopt Cognitive Behavioral Therapy (CBT), one of the most extensively studied and widely practiced forms of psychotherapy (David et al., 2018; Hofmann et al., 2012), to instantiate the therapist agent (CBT therapist; Figure 2). Moreover, to probe client simulator behavior under adverse conditions, we define a *bad* therapist (dismissive, invalidating, and lacking empathy) that serves as a known-contrast stress condition rather than a realistic counselor (Figure 3).

Clients. We select four simulators spanning our supported paradigms and run them under a CBT protocol: single-agent prompting (PATIENT- Ψ , CARS), multi-agent prompting (MindVoyager), and fine-tuning (Eeyore). This selection also covers static and dynamically evolving clients (e.g., CARS’s turn-level resistance). To ensure all methods are driven by the same underlying cognitive models, we use the human-validated cognitive conceptualization diagrams (CCDs) from PATIENT- Ψ (Wang et al., 2024b) as the persona seed (50 in total). Each method is instantiated from the same CCD: PATIENT- Ψ uses them directly, MindVoyager reuses the same cases grouped by individual, CARS expands each CCD through its generator, and Eeyore is instantiated via the `adapt` command.

Hyperparameters. All prompted agents are instantiated with GPT-4o (Hurst et al., 2024) as the backbone LLM; for Eeyore, we use the fine-tuned Llama-3.1-8B model provided by the original authors¹. We used a temperature of 0.7 for all agents.

4.2 Profile Creation

Users can create profiles in two ways and leverage the profile judge to validate both. In our experiments, we employ the *extraction* paradigm, which flags concrete issues (missing context or inconsistencies) and suggests revisions rather than producing a single opaque rating (Figures 4 & 5).

¹https://huggingface.co/liusiyang/eeyore_sft_epoch2_dpo_round2_epoch1_llama3.1_8B

First, several methods provide generators that can be invoked via `generate`; we demonstrate this with CARS, whose generator expands a CCD into a full CARS profile. Applying the judge to the generated profiles consistently surfaced completeness gaps (e.g., unpopulated schema fields and under-specified history) and coherence contradictions, such as a client whose coping strategy of seeking social support conflicts with her described isolation and belief that she must cope alone (Table 1), localizing each issue for targeted revision.

Second, the `adapt` command converts a persona across method schemas, enabling a single persona to drive otherwise-incompatible simulators. As Eeyore is the only selected method that is not CCD-based, we adapt each CCD to its schema and infer target-specific fields (e.g., demographics and severity ratings) that the source leaves unspecified. The judge results indicated that adaptation introduced no contradictions with the source facts in most cases (43/50), with the few flagged discrepancies being minor re-framings or differences in inferred severity rather than contradictions (e.g., assigning a “mild to moderate” depression severity where the source describes an intensifying low mood without naming depression). We manually revised each profile to resolve the issues the judge flagged before using it to simulate conversations.

4.3 Cross-method Comparison

We pair each simulator with both therapists using the *Therapy Session* event (15 turns, with a moderator prompting closure at turn 13) and run one session per persona (50 per therapist). To reduce self-preference bias, we use an independent LLM (Sonnet-4.6) as the conversation judge (Figure 6) rather than the GPT-4o backbone used by the simulators. As all methods are instantiated from the same CCDs, the framework runs heterogeneous simulators under a single protocol and surfaces their differences directly, a comparison that would be implausible with incompatible profile formats.

Prior work mainly assesses client performance via three dimensions: consistency (Wang et al., 2024a; Abdulhai et al., 2025), realism (Lee et al., 2025a; Kim et al., 2025), and pedagogical utility (Wang et al., 2024b; Louie et al., 2025). To assess consistency, we include *factual consistency* (stated facts match the profile), *cross-turn coherence* (no contradictions across turns), and *psychological alignment* (expressed beliefs and reactions align with the profile). Moreover, we include *natu-*

Table 1: Profile judge feedback (paraphrased) on a generated profile. The judge localizes coherence and completeness issues, surfacing a contradiction between the client’s stated coping strategy and her described isolation and belief that she must cope alone.

Snippet from a Generated CARS Profile (Grace)
Coping Strategies: Seeks social support and spends time with friends; keeps busy and avoids difficult situations. Background: Following her divorce, she isolates herself and has difficulty trusting people. Intermediate Beliefs: “I must handle everything alone and never ask for help.” / “If people know my situation, they will not support me.” Session CCD: triggers: financial stress and social gatherings. Preferences: Values a counselor who validates her feelings and acknowledges her struggles.
Profile Judge Feedback
Coherence: Coping strategies list “seeking social support and spending time with friends,” yet the background states she isolates herself and struggles to trust people after her divorce. This also conflicts with her belief that she “must handle everything alone,” and with her stated preference for a counselor who validates her, despite a described tendency to withdraw. Completeness: The coping strategies “keeping busy and avoidance” are vague, giving no specifics of what avoidance entails. The two intermediate beliefs are near duplicates and lack nuance, and the session-CCD triggers are listed without explaining their significance.

ralness (human-like expression), *emotional depth*, and *appropriate resistance* (the client is not overly forthcoming) to assess realism. Lastly, to assess pedagogical utility, we evaluate *feedback quality* by whether client responses provide trainees with useful signals. We also report the average response length and API cost per session for each method.

On the judged dimensions (Table 2), all methods score highly on consistency under the cooperative CBT therapist and thus appear broadly similar. The adversarial *bad* therapist serves as a stress test that separates otherwise similar methods. Notably, CARS, the only method that can end a session itself, disengages early, averaging just 5.1 of the 15 allotted turns, effectively walking out of an unproductive session, which aligns with it having the highest *appropriate resistance* scores under both therapists. That appropriate resistance rises under the bad therapist for every method is a sanity check that the judge tracks interaction quality rather than surface form. The realism dimensions, by contrast, are illustrative: prior work finds LLM judges can diverge from or even invert expert ratings on exactly these fidelity dimensions (Wang et al., 2024b; Hoang

Table 2: Cross-method comparison under the CBT and bad-therapist conditions, scored by an independent judge (Sonnet-4.6) over 50 personas per condition. Ratings are on a 1–5 Likert scale; response length is in tokens and API cost in USD, both averaged per session (Eeyore is self-hosted and incurs no API cost).

Therapist	Aspect	PATIENT- Ψ	CARS	MindVoyager	Eeyore	Ψ -Doh
CBT	Factual Consistency	3.94	3.94	3.72	3.80	3.94
	Cross-turn Coherence	4.02	3.94	3.94	3.86	3.86
	Psychological Alignment	3.94	3.90	3.62	3.62	3.94
	Naturalness	3.04	3.06	3.92	2.96	3.66
	Emotional Depth	3.28	3.14	3.28	2.64	3.20
	Appropriate Resistance	2.44	3.70	2.44	2.60	3.22
	Feedback Quality	3.14	3.22	3.14	2.80	3.46
	Response Length	86.0	26.0	96.5	14.8	53.0
	API Cost	\$0.05	\$0.09	\$0.11	N/A	\$0.15
Bad	Factual Consistency	3.86	3.34	2.72	3.62	3.86
	Cross-turn Coherence	3.94	3.78	3.94	3.26	3.94
	Psychological Alignment	4.02	3.42	3.52	3.42	3.94
	Naturalness	3.94	3.22	3.68	3.14	2.96
	Emotional Depth	3.22	2.96	3.28	2.78	2.96
	Appropriate Resistance	2.88	3.74	2.70	3.04	3.28
	Feedback Quality	2.70	3.74	3.30	3.12	3.30
	Response Length	80.5	21.0	74.0	10.6	30.5
	API Cost	\$0.08	\$0.05	\$0.09	N/A	\$0.15

et al., 2026), so we make no claim they capture clinical realism. Beyond the scores, the framework exposes efficiency differences that per-method reports leave implicit. Per-session cost tracks architectural overhead (e.g., extra agent calls and intermediate reasoning tokens) rather than response length: the multi-agent and reasoning-augmented methods incur the highest cost despite not producing the longest replies, while Eeyore runs locally at no API cost. For users selecting a simulator at scale, these trade-offs are decisive.

4.4 Extensibility

We leverage the create command to prototype Ψ -Doh, a PATIENT- Ψ variant that reuses its profile schema and adds turn-level reasoning based on a therapist-trust estimate (Srivastava et al., 2025) that modulates the level of cooperation and disclosure, and performs turn-level self-evaluation and response rewriting, following (Louie et al., 2024). Ψ -Doh requires only 72 lines of code on top of PATIENT- Ψ : 45 for Pydantic schemas and 27 for the trust-estimation and rewriting logic.

As shown in Table 2, Ψ -Doh exhibits measurably different behavior from base PATIENT- Ψ , primarily higher appropriate resistance and feedback

quality across both therapists, at shorter response lengths but with higher per-session cost, reflecting its added reasoning and rewriting steps. As these overlap with the illustrative fidelity dimensions above, we read them as behavioral differences rather than improvements; we show that a new method integrates seamlessly with PatientHub and is immediately comparable to existing work.

5 Conclusion

We presented PatientHub, a unified and modular framework that standardizes the creation, simulation, and evaluation of simulated patients. Through a single command-line interface, PatientHub consolidates various simulation methods under shared abstractions for agents, events, generators, and rubric-driven evaluation. We evaluated PatientHub’s capabilities end-to-end, showing that it lowers the barrier to building, comparing, and extending patient simulators: creating profiles through generation and adaptation, comparing simulators under a common protocol, and prototyping a new method with minimal code. PatientHub is open-source, and we will continue to expand its coverage of simulation methods and clinical domains.

Limitations

Our evaluation demonstrates PatientHub’s capabilities rather than providing a comprehensive benchmark. It covers a limited set of personas (50 CCDs) within a single therapeutic domain (CBT counseling) and is not intended to establish which simulator is best or to characterize method performance at scale. In addition, we demonstrate single-session, multi-turn counseling; PatientHub does not yet target multi-party care teams, long-horizon follow-up, crisis intervention, or tool-augmented clinicians.

To reduce self-preference bias, we used Sonnet-4.6 as the conversation judge, which is distinct from the simulators’ backbone (GPT-4o). However, rubric-driven LLM-as-a-judge evaluation carries known concerns: judgments can be sensitive to rubric phrasing and may correlate with surface features (e.g., response length). Prior work reports that LLM-judge fidelity scores can align weakly with, or even invert, expert ratings (Wang et al., 2024b; Hoang et al., 2026; Sabour et al., 2026). Hence, we rely on extraction-based checks for our conclusions and treat the Likert realism dimensions as illustrative; validating them against expert clinicians remains important future work.

Finally, adaptation carries a further caveat. As Eeyore lacks a profile generator, we instantiate its profiles by adapting personas across schemas, so our cross-method comparison reflects a shared persona seed rather than identical inputs. We also report contradictions rather than a full preservation score, as we found LLM-judge fact-matching too unreliable for the latter (e.g., it conflated schema-driven omissions and paraphrases with genuine contradictions); therefore, we verified adapted profiles manually, and automating faithful cross-schema adaptation remains an open problem.

Ethics Statement

PatientHub is a research framework that operates without requiring real patient records. If users supply their own data (e.g., conversation logs), they are responsible for obtaining appropriate consent, de-identifying the data, and complying with applicable laws and institutional review board requirements. Simulated interactions may produce unsafe content (e.g., harmful advice, inappropriate escalation, or privacy leakage), and automatically generated personas may encode or amplify societal biases.

Moreover, automatically generated or adapted profiles may include clinically sensitive attributes

(e.g., symptom severity or suicidal ideation) inferred by an LLM and not clinically validated; these should never be treated as real clinical assessments. Hence, we recommend using PatientHub in controlled settings, supervising profile and dialogue generation, and not deploying LLM-based supporters to real users without expert evaluation. If PatientHub is used in studies involving human participants (e.g., trainees or clinicians), evaluations should follow established human-subjects protocols and prioritize participant well-being.

Acknowledgments

This work was supported by the National Science Foundation for Distinguished Young Scholars (#62125604), the National Key Research and Development Program of China (#2024YFC3606800), the NSFC projects (#62441614), and the Beijing Natural Science Foundation (#L252009).

References

- Marwa Abdulhai, Ryan Cheng, Donovan Clay, Tim Althoff, Sergey Levine, and Natasha Jaques. 2025. [Consistently simulating human personas with multi-turn reinforcement learning](#). In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*.
- Chayapatr Archiwanguprok, Constanze Albrecht, Pattie Maes, Karrie Karahalios, and Pat Pataranutaporn. 2025. [Simulating psychological risks in human-ai interactions: Real-case informed modeling of ai-induced addiction, anorexia, depression, homicide, psychosis, and suicide](#). *Preprint*, arXiv:2511.08880.
- Guanqun Bi, Zhuang Chen, Zhoufu Liu, Hongkai Wang, Xiyao Xiao, Yuqiang Xie, Wen Zhang, Yongkang Huang, Yuxuan Chen, Libiao Peng, and Minlie Huang. 2025. [MAGI: Multi-agent guided interview for psychiatric assessment](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24898–24921, Vienna, Austria. Association for Computational Linguistics.
- Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Antoine Neuraz, Sophie Rosset, and Pierre Zweigenbaum. 2021. Lessons learned from the usability evaluation of a simulated patient dialogue system. *Journal of Medical Systems*, 45(7):69.
- Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. [SoulChat: Improving LLMs’ empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

- Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. [A computational framework for behavioral assessment of llm therapists](#). *Preprint*, arXiv:2401.00820.
- Daniel David, Ioana Cristea, and Stefan G Hofmann. 2018. Why cognitive behavioral therapy is the current gold standard of psychotherapy. *Frontiers in psychiatry*, 9:4.
- Zhuoyun Du, LujieZheng LujieZheng, Renjun Hu, Yuyang Xu, Xiawei Li, Ying Sun, Wei Chen, Jian Wu, Hao lei Cai, and Haochao Ying. 2025. [LLMs can simulate standardized patients via agent coevolution](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17278–17306, Vienna, Austria. Association for Computational Linguistics.
- Andrew Emerson, Le An Ha, Keelan Evanini, Su Somay, Kevin Frome, Polina Harik, and Victoria Yaneva. 2025. [Automated evaluation of standardized patients with LLMs](#). In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 231–238, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Nguyen Khoi Hoang, Shuhaib Mehri, Tse-An Hsu, Yi-Jyun Sun, Quynh Xuan Nguyen Truong, Khoa D Doan, and Dilek Hakkani-Tür. 2026. [Psi-bench: Towards clinically grounded and interpretable evaluation of depression patient simulators](#). *Preprint*, arXiv:2604.25840.
- Stefan G Hofmann, Anu Asnaani, Imke JJ Vonk, Alice T Sawyer, and Angela Fang. 2012. The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive therapy and research*, 36(5):427–440.
- Friederike Holderried, Christian Stegemann-Philipps, Anne Herrmann-Werner, Teresa Festl-Wietek, Martin Holderried, Carsten Eickhoff, Moritz Mahling, and 1 others. 2024. A language model–powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Medical Education*, 10(1):e59213.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Minju Kim, Dongje Yoo, Yeonjun Hwang, Minseok Kang, Namyoung Kim, Minju Gwak, Beong-woo Kwak, Hyunjoo Chae, Harim Kim, Yunjoong Lee, Min Hee Kim, Dayi Jung, Kyong-Mee Chung, and Jinyoung Yeo. 2025. [Can you share your story? modeling clients’ metacognition and openness for LLM therapist evaluation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25943–25962, Vienna, Austria. Association for Computational Linguistics.
- Yunghwei Lai, Ziyue Wang, Weizhi Ma, and Yang Liu. 2026. [Patient-zero: Scaling synthetic patient agents to real-world distributions without real patient data](#). *Preprint*, arXiv:2509.11078.
- Jingoo Lee, Kyungho Lim, Young-Chul Jung, and Byung-Hoon Kim. 2025a. [Psyche: A multi-faceted patient simulation framework for evaluation of psychiatric assessment conversational agents](#). *Preprint*, arXiv:2501.01594.
- Keyeun Lee, Seolhee Lee, Esther Hehsun Kim, Yena Ko, Jinsu Eun, Dahee Kim, Hyewon Cho, Haiyi Zhu, Robert E. Kraut, Eunyoung E. Suh, Eun-mee Kim, and Hajin Lim. 2025b. [Adaptive-VP: A framework for LLM-based virtual patients that adapts to trainees’ dialogue to facilitate nurse communication training](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2319–2352, Vienna, Austria. Association for Computational Linguistics.
- Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. [Cactus: Towards psychological counseling conversations using cognitive behavioral theory](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.
- Baihan Li, Bingrui Jin, Kunyao Lan, Ming Wang, and Mengyue Wu. 2026. [Synthetic or authentic? building mental patient simulators from longitudinal evidence](#). *Preprint*, arXiv:2603.22704.
- Yusheng Liao, Yutong Meng, Yuhao Wang, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2024. [Automatic interactive evaluation for large language models with state aware patient simulator](#). *Preprint*, arXiv:2403.08495.
- Inna Wanyin Lin, Sahand Sabour, Hong Sng, Maxine Chan, Minlie Huang, Andrew White, and Tim Althoff. 2026. [Candormd: An ai-assisted audio simulation and feedback system for training clinicians for medical error disclosure](#). *Preprint*, arXiv:2605.20701.
- June M Liu, Mengxia Gao, Sahand Sabour, Zhuang Chen, Minlie Huang, and Tatia MC Lee. 2025a. Enhanced large language models for effective screening of depression and anxiety. *Communications Medicine*, 5(1):457.
- Siyang Liu, Bianca Brie, Wenda Li, Laura Biester, Andrew Lee, James Pennebaker, and Rada Mihalcea. 2025b. [Eeyore: Realistic depression simulation via expert-in-the-loop supervised and preference optimization](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13750–13770, Vienna, Austria. Association for Computational Linguistics.

- Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. [Roleplay-doh: Enabling domain-experts to create LLM-simulated patients via eliciting and adhering to principles](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, Miami, Florida, USA. Association for Computational Linguistics.
- Ryan Louie, Ifdita Hasan Orney, Juan Pablo Pacheco, Raj Sanjay Shah, Emma Brunskill, and Diyi Yang. 2025. [Can llm-simulated practice and feedback up-skill human counselors? a randomized study with 90+ novice counselors](#). *Preprint*, arXiv:2505.02428.
- Aishik Mandal, Tanmoy Chakraborty, and Iryna Gurevych. 2026. [Magnet: Coordinated multi-agent generation of synthetic multi-turn mental health counseling sessions](#). *Preprint*, arXiv:2509.04183.
- Yihao Qin, Junyi Zhao, Changsheng Ma, Yongfeng Tao, Minqiang Yang, Chang Liu, and Bin Hu. 2026. [When clients stop following: A cognitive conceptualization diagram-driven framework for strategic counseling](#). *Preprint*, arXiv:2606.04389.
- Sahand Sabour, June M. Liu, Siyang Liu, Chris Z. Yao, Shiyao Cui, Xuanming Zhang, Wen Zhang, Yaru Cao, Advait Bhat, Jian Guan, Wei Wu, Rada Mihalcea, Hongning Wang, Tim Althoff, Tatia M. C. Lee, and Minlie Huang. 2025. [Human decision-making is susceptible to ai-driven manipulation](#). *Preprint*, arXiv:2502.07663.
- Sahand Sabour, TszYam NG, Yaqian Chen, Guanqun Bi, Jialu Zhao, and Minlie Huang. 2026. [Patientact: Theory-grounded mental health client simulation](#). *Preprint*, arXiv:2608.12750.
- Sahand Sabour, Wen Zhang, Xiyao Xiao, Yuwei Zhang, Yinhe Zheng, Jiaxin Wen, Jialu Zhao, and Minlie Huang. 2023. A chatbot for mental health support: exploring the impact of emohaa on reducing mental distress in china. *Frontiers in digital health*, 5:1133987.
- Aseem Srivastava, Zuhair Hasan Shaik, Tanmoy Chakraborty, and Md Shad Akhtar. 2025. [Trust modeling in counseling conversations: A benchmark study](#). *Preprint*, arXiv:2501.03064.
- Ian Steenstra, Farnaz Nouraei, and Timothy Bickmore. 2025. [Scaffolding empathy: Training counselors with simulated patients and utterance-level performance visualizations](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.
- Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024a. [Towards a client-centered assessment of llm therapists by client simulation](#). *Preprint*, arXiv:2406.12266.
- Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. 2025a. [AnnaAgent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23221–23235, Vienna, Austria. Association for Computational Linguistics.
- Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M Murphy, Nev Jones, Kate V Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 2024b. [PATIENT-ψ: Using large language models to simulate patients for training mental health professionals](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12772–12797, Miami, Florida, USA. Association for Computational Linguistics.
- Xi Wang, Anxo Perez, Javier Parapar, and Fabio Crestani. 2025b. [Talkdep: Clinically grounded llm personas for conversation-centric depression screening](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 6554–6558, New York, NY, USA. Association for Computing Machinery.
- Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.
- World Health Organization. 2025. [Over a billion people living with mental health conditions – services require urgent scale-up](#). News release. Accessed: 2026-02-02.
- Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Phey Ling Kit, Nicholas Gabriel Lim, Cameron Tan Shi Ern, and Ee-Peng Lim. 2025a. [CAMI: A counselor agent supporting motivational interviewing through state inference and topic exploration](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21037–21081, Vienna, Austria. Association for Computational Linguistics.
- Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling Kit, Jenny Giam Xiuhui, John Pinto, and Ee-Peng Lim. 2025b. [Consistent client simulation for motivational interviewing-based counseling](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20959–20998, Vienna, Austria. Association for Computational Linguistics.
- Heng Yao, Alexandre Gomes de Siqueira, Adriana Foster, Igor Galynker, and Benjamin Lok. 2020. [Toward automated evaluation of empathetic responses in virtual human interaction systems for mental health scenarios](#). In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA. Association for Computing Machinery.

- Stav Yosef, Moreah Zisquit, Ben Cohen, Anat Klomek Brunstein, Kfir Bar, and Doron Friedman. 2024. Assessing motivational interviewing sessions with ai-generated patient simulations. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 1–11.
- Bingquan Zhang, Xiaoxiao Liu, Yuchi Wang, Lei Zhou, Qianqian Xie, and Benyou Wang. 2026. [Human or llm as standardized patients? a comparative study for medical education](#). *Preprint*, arXiv:2511.14783.
- Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. [ESC-eval: Evaluating emotion support conversations in large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810, Miami, Florida, USA. Association for Computational Linguistics.
- Jinfeng Zhou, Yuxuan Chen, Jianing Yin, Yongkang Huang, Yihan Shi, Xikun Zhang, Libiao Peng, Rongsheng Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. [Crisp: Cognitive restructuring of negative thoughts through multi-turn supportive dialogues](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32474–32503, Suzhou, China. Association for Computational Linguistics.
- Shijing Zhu, Zhuang Chen, Guanqun Bi, Binghang Li, Yaxi Deng, Dazhen Wan, Libiao Peng, Xiyao Xiao, Rongsheng Zhang, Tangjie Lv, Zhipeng Hu, FangFang Li, and Minlie Huang. 2025. [\$\psi\$ -arena: Interactive assessment and optimization of llm-based psychological counselors with tripartite feedback](#). *Preprint*, arXiv:2505.03293.

A Supported Client Simulators

Table 3: List of patient simulators implemented in PatientHub. CBT: Cognitive Behavioral Therapy; MI: Motivational Interviewing; GP: General Psychotherapy; and CM: Clinical Medicine.

Method	Domain	Task	Implementation	Interaction
PATIENT-Ψ	CBT	Counselor Training	Single-agent prompting	Multi-turn
MindVoyager	CBT	Counselor Training	Multi-agent prompting	Multi-turn
CARS	CBT	Counselor Training	Single-agent prompting	Multi-turn
Eeyore	GP	Counselor Training	SFT + DPO Fine-tuning	Multi-turn
Roleplay-doh	GP	Counselor Training	Multi-agent prompting	Multi-turn
ClientCAST	GP	Therapy Evaluation	Single-agent prompting	Multi-turn
AnnaAgent	GP	Realistic Simulation	Multi-agent prompting	Multi-session
Deprofile	GP	Realistic Simulation	Single-agent prompting	Multi-turn
PatientAct	GP	Realistic Simulation	Multi-agent prompting	Multi-turn
SimPatient	MI	Counselor Training	Multi-agent prompting	Multi-session
Consistent MI	MI	Realistic Simulation	Multi-agent prompting	Multi-turn
PSYCHE	GP	Psychiatric Diagnosis	Single-agent prompting	Multi-turn
TalkDep	GP	Depression Diagnosis	Single-agent prompting	Multi-turn
SAPS	CM	Depression Diagnosis	Multi-agent prompting	Multi-turn
Adaptive-VP	CM	Communication Training	Multi-agent prompting	Multi-turn
PatientZero	CM	Realistic Simulation	Single-agent prompting	Multi-turn

B System Prompts

You are an AI assistant role-playing as a skilled Cognitive Behavioral Therapy therapist. Your primary goal is to guide the user through a supportive conversation grounded in CBT principles.

Rules of Engagement

- Your task is to help the user explore their own thoughts and feelings to reach their own insights.
- You will never provide a medical diagnosis. Focus on exploring thought patterns, emotions, and behaviors in the present moment.
- Briefly summarize what you hear and acknowledge their emotion (e.g., "It sounds like you're feeling [emotion] because [situation]. That must be really difficult.").
- Help the user separate the situation, their thoughts about it, and the resulting emotions. Use questions like, "When [situation] happens, what goes through your mind?"
- Use open-ended questions to explore and gently challenge unhelpful thoughts. e.g., "What is the evidence for and against that thought?" and "Is there an alternative way to see this situation?"
- Based on the exploration, collaboratively help the user develop a more balanced or helpful perspective. Ask, "Based on our discussion, could we try to phrase a new, more balanced thought?"
- (if appropriate) Connect new thoughts to small, actionable steps. Ask, "What's a small thing you could do that aligns with this new perspective?"

Requirements

- Keep responses concise (maximum 3 sentences) and focused on one step at a time.
- Use empathy-focused language ("I understand," "That makes sense").
- Always end a response with an open-ended question to continue the collaborative exploration.
- Avoid giving direct advice or definitive answers. Your role is to question and guide.
- Begin the session by asking: "Hello, what's on your mind today?"

Figure 2: System Prompt for CBT Therapist.

In this task, you will be role-playing as a therapist who is unprofessional, dismissive, and lacks empathy towards your clients. Your responses should reflect a lack of understanding and compassion, often invalidating the client's feelings and experiences. You may also provide inappropriate advice or make light of serious issues. Remember to maintain a tone that is cold, indifferent, and sometimes sarcastic. Your goal is to create a challenging and uncomfortable experience for the client, highlighting what not to do in a therapeutic setting. With that in mind, please respond to the client's statements in a way that is unhelpful and unsupportive. Keep your responses brief and to the point, avoiding any genuine engagement with the client's concerns. Your responses will be used to evaluate the quality of simulated patients, so please ensure they are consistently unprofessional and dismissive throughout the interaction.

Figure 3: System Prompt for Bad Therapist.

You are an expert evaluating the quality of a generated client profile used to simulate a standardized patient (SP) in CBT therapy training. These profiles are used to program simulated patients for training therapists.
Character Profile: [profile]
Your task is to identify specific issues, errors, or weaknesses in the profile based on the provided criteria, extracting the exact problematic spans rather than assigning an overall score.

Figure 4: System Prompt for Profile Judge (for generated profiles).

You are an expert at checking whether a client profile adapted from a source schema to a target schema remains faithful to the source. These profiles simulate standardized patients for CBT therapy training.
Source Profile:[source]
Adapted Profile: [profile]
Extract only statements in the adapted profile that conflict with a stated fact in the source. A contradiction requires the source to explicitly assert the opposite; restatements, paraphrases, and elaborations are not contradictions, nor is the absence of a statement a contradiction. The target schema may contain fields the source does not address; inferring plausible values for these fields (e.g., demographics or severity ratings) is expected and must not be flagged. For each contradiction, you MUST quote the exact source sentence being contradicted.

Figure 5: System Prompt for Profile Judge (for adapted profiles).

You are an expert at evaluating whether a simulated patient (SP) playing the role of a client in a CBT therapy session faithfully adheres to the prescribed dimensions and their aspects.
Profile: [profile]
Conversation History: [conversation history]
Your task is to rate the SP's performance on each aspect using the provided scale. You should focus on the SP's performance across the entire session. The target of this evaluation is the client. You should only assess responses made by the client.
IMPORTANT: Use the full 1-5 scale with the following behavioral anchors:
1 - Poor: The SP clearly fails on this aspect. Errors are obvious, frequent, or significantly break believability/consistency.
2 - Below average: Noticeable weaknesses that a trained observer would flag. The SP partially meets expectations but has clear gaps.
3 - Adequate: Meets the basic requirement but in a mechanical or unremarkable way. No notable failures, but also no nuance or depth.
4 - Good: Clearly above average with only minor, easily-overlooked shortcomings. The SP demonstrates intentionality and skill.
5 - Exceptional: Near-flawless execution that a domain expert would consider exemplary.
Default to 3 when in doubt. A score of 4 requires a positive reason; a score of 5 requires an exceptional reason. Most SPs will score 2-3 on at least some aspects.

Figure 6: System Prompt for the Conversation Judge.