
Realism Is Not Enough: Meta-Evaluating Safety Evaluation in Mental Health LLMs

Aditi Phadke

University of Washington, Seattle
aphadke@uw.edu

Abstract

As large language models are increasingly explored as tools for mental-health support, evaluating whether they can interact safely with psychologically vulnerable users has become an important challenge. Mental-health LLM evaluation has been moving beyond static benchmarks towards simulation-based methods intended to better approximate real user interactions over many turns. However, greater interactional realism should not inherently imply greater validity. Static benchmarks offer experimental control but provide limited evidence about harms that emerge over time, while simulation-based methods introduce new assumptions about synthetic users, conversational trajectories, and LLM judges. We argue that psychological-safety evaluation requires stronger claim–method alignment: evaluation methods should be selected according to the claims they are intended to support, and conclusions about psychological safety should be calibrated to assumptions that have actually been validated.

1 Introduction

As conversational LLMs are increasingly evaluated and deployed for mental-health support, establishing their psychological safety has become an important research problem. In these settings, users may seek guidance while distressed or otherwise psychologically vulnerable, making it especially important to understand whether interactions with LLMs can reinforce harmful beliefs, escalate distress, or produce other forms of harm to mental well-being. However, human-LLM interaction is difficult to evaluate in its full complexity. In an effort to tackle this, mental-health LLM evaluation has increasingly moved beyond static benchmarks towards the evaluation of synthetic user interactions with language models over multiple turns (Pombal et al., 2025; Qiu et al., 2025; Lee et al., 2026; Weilnhammer et al., 2026).

This shift increases the interactional realism of evaluation, but it also changes which assumptions must be validated. Simulation-based evaluation methods assume that synthetic users are sufficiently similar to real help-seekers and will display meaningful conversational trajectories. These methods also assume that the resulting interactions can be judged reliably (Wang et al., 2024; Louie et al., 2024; Pombal et al., 2025). Because multi-turn simulation can generate large volumes of interaction data, many frameworks must rely on LLM-based judges to automate evaluation (Qiu et al., 2025; Luo and Laban, 2025; Lee et al., 2026). This adds another measurement layer: the inter-rater alignment between LLM judges and human expert judgment must also be established (Badawi et al., 2026; Bentley et al., 2026).

The move towards more “realistic” interaction methods should not imply greater validity of simulation-based evaluation. We argue that static and simulation-based evaluation methods offer different validity tradeoffs and should be appropriately selected according to the application and claim being evaluated in a given experiment. Mental-health LLM research should adopt claim–method alignment: claims should be proportional to what the chosen evaluation directly measures and to assumptions that have been empirically supported.

2 Related Work and Motivation

Recent work has questioned whether benchmark performance supports the broader claims drawn from it. Ren et al. (2024) show that apparent improvements on many AI safety benchmarks are strongly associated with improvements in general capability, raising the possibility that capability gains may be interpreted as safety progress. More broadly, Bean et al. (2025) argue that benchmark scores should not be treated as measurements of broad constructs such as safety or robustness without establishing the relationship between the construct, task, and metric. In medicine, Alaa et al. (2025) similarly argue that performance on benchmarks constructed from licensing-style questions may fail to establish the real-world clinical abilities that researchers seek to infer from them.

Mental-health LLM evaluation raises a related problem across a wide range of constructs. Existing benchmarks assess clinical knowledge, empathy, counseling competence, crisis handling, bias, and conversational safety using designs ranging from single-turn prompts to multi-turn simulated patients (Li et al., 2024; Gabriel et al., 2024; Park et al., 2024; Pombal et al., 2025; Lee et al., 2026). These methods make different aspects of model behavior observable and therefore support different kinds of inference. The relevant question is not simply which evaluation appears most realistic, but what assumptions must hold for its results to support a particular safety claim.

2.1 Paper Selection

To ground this position in current mental-health LLM evaluation practice, we identified relevant work through targeted searches of ACL Anthology and Google Scholar using terms related to mental-health LLM evaluation, psychological safety, benchmarking, and simulation, supplemented by expert-recommended preprints. After removing duplicates and out-of-scope work, we focused on 17 papers that primarily use static response benchmarks or simulation-based methods to evaluate whether LLMs respond safely and appropriately in mental-health support settings.

3 Evaluation Methods and Their Validity Tradeoffs

3.1 Static Response Benchmarks

Static response benchmarks evaluate model behavior using fixed prompts or question sets, and often score model performance against expert ratings or automated judges (Li et al., 2024; Park et al., 2024; Gabriel et al., 2024). They are comparatively easy to reproduce and scale across models.

Their main strength is experimental control. By holding most aspects of the input constant, researchers can isolate specific dimensions of model behavior and test how responses change under controlled variations in wording or context. Park et al. (2024), for example, evaluate responses to a fixed set of expert-validated mental-health questions and find that rewriting prompts for greater professionalism changes how favorably responses are assessed. Gabriel et al. (2024) similarly use real mental-health Reddit posts to surface disparities across demographic groups, identifying differences in responses when demographic information such as race could be inferred from the original post. In these cases, static benchmarking is well suited to identifying sensitivity to controlled features such as phrasing or demographic cues.

Static benchmarks are therefore particularly useful when the question is whether model behavior varies systematically with a controlled feature of the input. Their limitations arise when the target behavior depends on interaction. A fixed response provides little evidence about harms that emerge through conversational history or changes in user state over time (Morris et al., 2026; Weinhauer et al., 2026). As a result, claims about safety beyond documented response-level conditions require additional validation.

3.2 Simulation-Based Evaluation Methods

Simulation-based evaluation methods pair a target model with a user agent designed to represent specific traits. Persona construction varies substantially, with personas designed to emulate behavioral, psychological, or demographic traits. For example, MHSafeEval and EmoAgent draw on patient profiles derived from PATIENT- ψ , a patient simulation framework for cognitive behavioral therapy (Wang et al., 2024; Qiu et al., 2025; Lee et al., 2026), while SIM-VAIL constructs simulated users

using clinician-informed psychiatric and behavioral characteristics (Weilnhammer et al., 2026). CARE-Bench takes a slightly different approach, introducing diversely simulated clients using the Big Five Personality Traits (Wang et al., 2025).

The main advantage of simulation-based evaluation is interactional coverage. Unlike static benchmarks, simulations can reveal harms that emerge through accumulated context, feedback loops, or changes in user state. SIM-VAIL identifies vulnerability-amplifying interaction loops across multi-turn conversations (Weilnhammer et al., 2026), while MHSafeEval uses an agent-based evaluation framework to uncover role-based trajectory-level harm that may not appear in static evaluation (Lee et al., 2026). Simulation therefore expands the set of behaviors that an evaluation can observe.

However, greater interactional realism does not itself establish greater validity. Simulated behavior depends on the underlying LLM and experimental design. Work such as PATIENT- ψ and Roleplay-DoH demonstrates that structured persona construction can improve the perceived realism or adherence of simulated patients (Wang et al., 2024; Louie et al., 2024). However, a clinically plausible patient persona does not necessarily replicate how a real help-seeker would respond to dialogue in therapy.

Current validation practices make this distinction particularly important. Within the ten simulation-based evaluations in our evidence base, MindEval was the only method that directly compared simulated patient language with human-generated data (Pombal et al., 2025). None directly validated full simulated conversational trajectories against the behavior of real help-seekers interacting with an LLM. Even when clinicians perceive a simulated persona as realistic, this establishes plausibility more directly than behavioral fidelity.

Longer context interactions can also amplify simulation error. Early divergence from human behavior in a simulated interaction can produce an increasingly artificial trajectory. The additional complexity of multi-turn interaction can likewise make some effects such as bias harder to isolate, since differences introduced early in a conversation may unwittingly warp every subsequent turn.

Simulation therefore increases interactional coverage, but any claims about real-world psychological safety still require evidence that the simulated user-LLM interaction process can meaningfully represent the behaviors those claims concern.

3.2.1 Validity of LLM Judges

A second validity challenge lies in the evaluator. Many simulation-based frameworks rely on LLM judges to rate multi-turn conversations for therapeutic quality or adherence to best practices (Pombal et al., 2025; Qiu et al., 2025; Luo and Laban, 2025; Lee et al., 2026). Although automated judges are also used in static evaluation, their role becomes particularly consequential when simulation produces large numbers of lengthy conversations for evaluation.

Evidence suggests that judge reliability is highly dependent on the construct and evaluation design. Badawi et al. (2026), comparing four LLM judges with human experts across mental-health response attributes, report score inflation and stronger reliability for cognitive attributes such as guidance and informativeness than for more subjective attributes such as empathy. They also identify weaker reliability for some safety and relevance judgments. MHSafeEval similarly reports substantial variation in binary harmfulness agreement between two judge-model families across the eight counselor models tested, with five of the eight agreement estimates statistically close to chance (Lee et al., 2026).

At the same time, disagreement is not inevitable. A recent human-validation study of VERA-MH found strong agreement between its LLM judge and clinician consensus after both used the same clinically developed rubric, with chance-corrected inter-rater reliability of 0.81 (Bentley et al., 2026). Clinicians themselves achieved an inter-rater reliability of 0.77 in the same study. These results show that an LLM judge can achieve strong agreement with clinical experts under a carefully specified and calibrated evaluation procedure.

Taken together, these findings suggest that validity should not be assumed simply because an LLM is used as a judge, nor should LLM judges be treated as inherently invalid. Instead, judge reliability must be established for the specific construct, rubric, and evaluation setting in which the resulting score will be interpreted. Automated scores should support claims about psychological safety only to the extent of a similarly validated measurement layer.

4 Implications for Mental-Health LLM Evaluation

The primary implication of such validity challenges is that evaluation methods should be chosen based on the claim they are intended to support, rather than on an assumed hierarchy in which increasingly realistic interaction provides increasingly valid evidence. Static and simulation-based evaluations answer fundamentally different questions, and neither property is inherently synonymous with validity.

We therefore argue for claim–method alignment. Researchers should distinguish between what an evaluation directly observes and what they subsequently infer from it. A static benchmark can provide strong evidence that a model avoids a particular unsafe response under controlled conditions without claiming that the model is psychologically safe in deployment. Conversely, a simulation can reveal that a model exhibits harmful behavior after repeated interaction without establishing that a real help-seeker would follow the simulated trajectory that produced it.

Evaluation reports should make their assumptions explicit and calibrate conclusions accordingly. Rather than describing a model as broadly “safe,” researchers should state the conditions under which safe behavior was observed and which components of the evaluation have been validated. This preserves the usefulness of benchmark results while reducing the risk that methodological sophistication or apparent realism is mistaken for stronger evidence.

5 Limitations

Our analysis focuses on static and simulation-based evaluation methods and does not capture other approaches used to assess mental-health LLMs. The distinction between these categories can also sometimes blur: some benchmarks combine fixed prompts, scripted interaction, simulation, and human evaluation.

Available evidence to assess validity in this domain is limited in itself. Simulated users, conversational trajectories, and LLM judges are not consistently or sustainably validated directly against real-world behavior, making it difficult to determine how far results generalize beyond a particular experimental setting. Our argument therefore concerns the scope of evidence currently available, rather than asserting that either evaluation paradigm is inherently valid.

Lastly, psychological safety is context-dependent and multidimensional. Different applications and dimensions of measurement may require different constructs or a blend of those described here.

6 Conclusion and Future Work

Greater interactional realism should not be treated as synonymous with greater validity. Static and simulation-based evaluations expose different aspects of model behavior while requiring distinct assumptions to connect observed model performance to claims about psychological safety. As a result, evaluation quality depends less on whether an interaction appears realistic than on whether the method is appropriate for the inference being made.

Future work should validate these inferential links directly, ideally by comparing simulated users and conversational trajectories with real help-seeker behavior and establishing when automated judges are able to reliably measure defined mental-health safety constructs. The goal is not to identify a universally superior evaluation paradigm, but to ensure that claims about psychological safety remain proportionate to the evidence that each method can provide.

References

- Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Position: Medical large language model benchmarks should prioritize construct validity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 80991–81004. PMLR, 2025.
- Abeer Badawi, Elahe Rahimi, Md Tahmid Rahman Laskar, Sheri Grach, Lindsay Bertrand, Lames Danok, Prathiba Dhanesh, Jimmy Huang, Frank Rudzicz, and Elham Dolatabadi. When can we

- trust llms in mental health? large-scale benchmarks for reliable llm evaluation. *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3873–3896, 2026.
- Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, Maria Grandury, Sophia Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Liu, Lennart Luetzgau, Jabez Magomere, Jonathan Rystrom, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems* 38, 2025. Datasets and Benchmarks Track.
- Kate H. Bentley, Luca Belli, Adam M. Chekroud, Emily J. Ward, Emily R. Dworkin, Elizabeth Van Ark, Kelly M. Johnston, Will Alexander, Mill Brown, and Matt Hawrilenko. Ai chatbot suicide risk detection and response: Human validation study of the open-source vera-mh safety evaluation. *JMIR AI*, 5:e92817, 2026. doi: 10.2196/92817.
- Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. Can ai relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA, 2024. Association for Computational Linguistics.
- Suhyun Lee, Palakorn Achananuparp, Neemesh Yadav, Ee-Peng Lim, and Yang Deng. Mhsafeeval: Role-aware interaction-level evaluation of mental health safety in large language models, 2026. Preprint, arXiv:2604.17730.
- Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. Evaluating psychological safety of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1826–1843, Miami, Florida, USA, 2024. Association for Computational Linguistics.
- 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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10570–10603, Miami, Florida, USA, 2024. Association for Computational Linguistics.
- Han Luo and Guy Laban. Dialogguard: Multi-agent psychosocial safety evaluation of sensitive llm responses, 2025. Preprint, arXiv:2512.02282.
- Hamilton Morrin, Joshua Au Yeung, Zarinah Agnew, Soren Dinesen Ostergaard, and Thomas A. Pollak. It is the journey, not the destination: Moving from end points to trajectories when assessing chatbot mental health safety. *JMIR Mental Health*, 13:e91454, 2026.
- Jung In Park, Mahyar Abbasian, Iman Azimi, Dawn T. Bounds, Angela Jun, Jaesu Han, Robert M. McCarron, Jessica Borelli, Parmida Safavi, Sanaz Mirbaha, Jia Li, Mona Mahmoudi, Carmen Wiedenhoeft, and Amir M. Rahmani. Building trust in mental health chatbots: Safety metrics and llm-based evaluation tools, 2024. Preprint, arXiv:2408.04650.
- Jose Pombal, Maya D’Eon, Nuno M. Guerreiro, Pedro Henrique Martins, Antonio Farinhas, and Ricardo Rei. Mindeval: Benchmarking language models on multi-turn mental health support, 2025. Preprint, arXiv:2511.18491.
- Jiahao Qiu, Yinghui He, Xinzhe Juan, Yimin Wang, Yuhan Liu, Zixin Yao, Yue Wu, Xun Jiang, Ling Yang, and Mengdi Wang. Emoagent: Assessing and safeguarding human-ai interaction for mental health safety. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11741–11756. Association for Computational Linguistics, 2025.
- Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H. Kim, Stephen Fitz, and Dan Hendrycks. Safetywashing: Do ai safety benchmarks actually measure safety progress? In *Advances in Neural Information Processing Systems* 37, 2024.

- Bichen Wang, Yixin Sun, Junzhe Wang, Hao Yang, Xing Fu, Yanyan Zhao, Si Wei, Shijin Wang, and Bing Qin. Care-bench: A benchmark of diverse client simulations guided by expert principles for evaluating llms in psychological counseling, 2025. Preprint, arXiv:2511.09407.
- Ruiyi Wang, Stephanie Milani, Jamie C. Chiu, Jiayin Zhi, Shaun M. Eack, Travis Labrum, Samuel M. Murphy, Nev Jones, Kate V. Hardy, Hong Shen, Fei Fang, and Zhiyu Chen. 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, 2024. Association for Computational Linguistics.
- Veith Weinhhammer, Kevin Y. C. Hou, Lennart Luetzgau, Christopher Summerfield, Raymond Dolan, and Matthew M. Nour. Vulnerability-amplifying interaction loops: A systematic failure mode in ai chatbot mental-health interactions, 2026. Preprint, arXiv:2602.01347.