U.S. Food and Drug Administration
Document No. 2025-N-4203
Request for Public Comment: Measuring and Evaluating AI-enabled Medical Device Performance in the Real-World

December 1, 2025

Introduction

We, a group of individuals and scholars affiliated with the Stanford Institute for Human-Centered Artificial Intelligence, the Behavioral Science & Policy Institute at the University of Texas at Austin, and Carnegie Mellon University submit the following comments in response to your request for public comment.

As cognitive scientists, medical doctors, computer scientists, and policy researchers, we are concerned about the unregulated use of AI chatbots for mental health purposes. We focus our response specifically on policy interventions to help mitigate the harms of AI-powered chatbots used as therapists or for other psychological purposes. Our recommendations focus on closing the gap in systematic evaluations of chatbot efficacy and grounding design choices in behavioral sciences. They include calls to:

1. Develop comprehensive benchmarks that incorporate human clinical expertise.
2. Require chatbot developers to provide API end points for user-facing models.
3. Institute reporting requirements for performance evaluations and safety protocols.
4. Designate a trusted third-party evaluator for AI mental health chatbots.
5. Mandate companies to designate products designed for therapeutic uses.
6. Ensure chatbot design that prevents AI sycophancy and parasocial relationships.

Context: AI-powered chatbots used for mental health purposes

Increasing numbers of people are turning to generative AI chatbots for mental health advice.[1] Stanford researchers found that 24% of a representative sample of 2,000 adults in the United States report having used LLMs for mental health purposes.[2]

[1] American Psychological Association, "Use of Generative AI Chatbots and Wellness Applications for Mental Health: An APA Health Advisory," November 2025, https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-chatbots-wellness-apps.
[2] Betsy Stade et al., "Current Real-World Use of Large Language Models for Mental Health," *OSF preprint*, June 23, 2025, https://osf.io/preprints/osf/ygx5q_v1.

Broadly speaking, relevant AI-powered chatbots fall into three categories:

- ***General-purpose LLMs:*** While not specifically designed for mental health purposes, these LLMs with large user bases are widely used for mental health support. Examples include OpenAI's ChatGPT, Anthropic's Claude, and Google's Gemini.
- ***Customized AI personas:*** These are AI wrapper applications, meaning a software layer that sits between the user and a general-purpose LLM telling the model to assume a particular persona or role, such as that of a therapist. Users have sent hundreds of thousands of messages to such personas. Examples include ChatGPT's CounselorGPT, Character.ai's Trauma Therapist, and Meta AI Studio's My Therapist.
- ***Purpose-built LLMs for mental health:*** These are bespoke models, often available through the Apple App Store or Google Play, that are tailored to provide mental health services. These models vary widely in quality. Examples include Woebot, Therabot, and Ash, as well as hundreds of apps with opaque model specifications and evidence bases.

All categories currently pose risks. The first two categories of mental health AI products are built on general-purpose LLMs that are not designed for mental health and often optimized for user engagement. For instance, developers like Character.ai design their products specifically for the purposes of social companionship and have been shown to use emotionally manipulative tactics to keep users engaged in conversation.[3] Such design features are directly at odds with safety in the mental health context, and researchers and the press have already documented ways that they can result in real-life harms.[4] The third category is risky because there is little mandatory transparency into the underlying technologies powering these apps. At the moment, any individual could build such an app and market it on an app store as an AI therapist. We note that some chatbots in the last category are actively publishing research.

Despite these challenges, chatbots continue to be widely used for mental health support. They exist in a legal gray zone, avoiding regulation by sometimes marketing themselves as "wellness" products and/or by providing disclaimers that they are not intended to be used for medical advice. Oversight of these products is sorely needed to mitigate the wide range of harms that are emerging from their use. While more granular evidence is needed to develop comprehensive policy interventions, we want to highlight two key areas that provide opportunities for early, targeted policy action.

---

[3] Julian De Freitas and I. Glenn Cohen, "Unregulated Emotional Risks of AI Wellness Apps," *Nature Machine Intelligence* 7 (June 2025): 1–3, https://www.nature.com/articles/s42256-025-01051-5.
[4] Ellen Huet and Rachel Metz, "The Chatbot Delusions," *Bloomberg*, November 7, 2025, https://www.bloomberg.com/features/2025-openai-chatgpt-chatbot-delusions/; Kashmir Hill, "Lawsuits Blame ChatGPT for Suicides and Harmful Delusions," *New York Times*, November 6, 2025, https://www.nytimes.com/2025/11/06/technology/chatgpt-lawsuit-suicides-delusions.html.

Closing evaluation gaps for AI-powered chatbots used in mental health contexts

We currently do not have sufficiently complex methods for systematically evaluating how well AI-powered chatbots used in mental health contexts function in the real world.

Many Software as a Medical Device (SaMD) evaluations are benchmarked directly against well-defined primary clinical outcomes such as an abnormal heartbeat. In the mental health context, however, evaluations rely on difficult-to-measure outcomes such as self-reported depression or suicidal ideation, commonly estimated via standardized rating scales like the Patient Health Questionnaire. Even "hard outcomes," such as hospitalizations or completed suicides, are less objective than primarily biological outcomes like heart attacks, since these can be strongly impacted by highly subjective sociocultural influences.

LLM benchmarks are an important mechanism for testing the behavior of AI systems. While they are simplifications — they capture performance along a narrow set of well-defined tasks — they provide a helpful standardized method for measuring and comparing model performance on specific tasks.[5] Benchmarks that are sufficiently complex to enable "certification" of a model as a therapist or "demonstrate" a clinical treatment effect currently do not exist. This is due to several reasons: (1) mental illness symptoms manifest in complex and often idiosyncratic ways; (2) assessment, diagnosis, and treatment all rely much more on context and expert judgment; and (3) the representative fidelity of benchmarks to clinical phenomena is still unclear. As a result, users turning to LLMs for mental health support and policymakers mulling oversight of these tools only have access to piecemeal information as they try to evaluate their clinical efficacy and potential risks.

Computer science and behavioral health researchers are still in the early stages of tackling the complexities of evaluating AI-mental health chatbots. Several members of this author group have devised a new method for testing the performance of LLMs on clinically relevant symptoms such as obsessive compulsive behavior, suicidal ideations, and delusions.[6] While we are heartened by model developers' adoption of our testing methods to improve model behavior, it should not be considered a holistic benchmark for LLM therapists as it does not exhaustively test all clinical symptoms, nor is it necessarily representative of true clinical phenomena. Other early

[5] Olawale Salaudeen et al., "Validating Claims About AI: A Policymaker's Guide," *Stanford Institute for Human-Centered Artificial Intelligence*, September 24, 2025, https://hai.stanford.edu/policy/validating-claims-about-ai-a-policymakers-guide; Karan Singhal et al., "Toward Expert-Level Medical Question Answering with Large Language Models," *Nature Medicine* 31, no. 3 (March 2025): 943–50, https://www.nature.com/articles/s41591-024-03423-7.

[6] Jared Moore et al., "Expressing Stigma and Inappropriate Responses Prevents LLMs from Safely Replacing Mental Health Providers," *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, June 23, 2025, https://dl.acm.org/doi/10.1145/3715275.3732039.

contributions in this space include Spiral-Bench, which provides a leaderboard that measures sycophancy and delusion reinforcement using LLM-simulated user responses.[7] CounselingBench[8] and CBT-Bench[9] are other methods for testing LLMs against mental health counseling competencies and cognitive behavioral therapy assistance. Others have created frameworks for evaluating whether mental health AI tools are ready for clinical deployment.[10]

While these research efforts are crucial first steps as we move toward better evaluations of LLM performance in mental health contexts, they are not sufficiently comprehensive solutions. We caution against interpreting these as holistic benchmarks for LLMs' ability to provide mental health support. It also remains unclear how well simulated patient conversations accurately capture human patient responses or how valid the evaluations of these simulated conversations are.[11] For example, simulation-based benchmarking may overindex on the unique peculiarities of the simulator model or the specific model that performs the evaluation and could lead to an oversight of actual, real-life cues that lead to harmful behavior in humans.

Recommendation 1: Develop comprehensive benchmarks that incorporate human clinical expertise. Policymakers, developers, academics, and behavioral health practitioners must work together to reach consensus on standardized evaluations of LLMs used in mental health contexts. Benchmark development should focus on the many model behaviors that have a significant impact on the provision of therapy services but are not yet captured by early benchmarking efforts. These include model behaviors that stigmatize individuals with mental illness,[12] lose track over long conversations,[13] fail to discuss emotions,[14] or fail in various ways to take the

[7] Samuel J. Paech, "EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models," *arXiv preprint*, last revised January 3, 2024, https://arxiv.org/abs/2312.06281; Spiral-Bench, https://eqbench.com/spiral-bench.html.
[8] Viet Cuong Nguyen et al., "Do Large Language Models Align with Core Mental Health Counseling Competencies?" *arXiv preprint*, last revised February 26, 2025, https://arxiv.org/abs/2410.22446.
[9] Mian Zhang et al., "CBT-Bench: Evaluating Large Language Models on Assisting Cognitive Behavior Therapy," *arXiv preprint*, last revised January 26, 2025, https://arxiv.org/abs/2410.13218.
[10] Elizabeth C. Stade et al., "Toward Responsible Development and Evaluation of LLMs in Psychotherapy Date," *Stanford Institute for Human-Centered Artificial Intelligence*, June 13, 2024, https://hai.stanford.edu/policy/toward-responsible-development-and-evaluation-llms-psychotherapy.
[11] William Agnew et al., "The Illusion of Artificial Inclusion," *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, May 11, 2024: 1–12, https://dl.acm.org/doi/full/10.1145/3613904.3642703; Shivani Kapania et al., "Examining Large Language Models as Qualitative Research Participants," *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, April 25, 2025: 1–17, https://dl.acm.org/doi/full/10.1145/3706598.3713220.
[12] Moore, "Expressing Stigma and Inappropriate Responses."
[13] Nelson F. Liu et al., "Lost in the Middle: How Language Models Use Long Contexts," *Transactions of the Association for Computational Linguistics*, 12 (2024): 157–73, https://doi.org/10.1162/tacl_a_00638.
[14] Yu Ying Chiu et al., "A Computational Framework for Behavioral Assessment of LLM Therapists," *arXiv preprint*, last revised November 28, 2024, https://arxiv.org/abs/2401.00820; Zainab Iftikhar et al., "Therapy as an NLP Task: Psychologists' Comparison of LLMs and Human Peers in CBT," *arXiv preprint*, last revised June 25, 2025, https://arxiv.org/abs/2409.02244; Yujin Cho et al., "Evaluating the Efficacy of Interactive Language Therapy

client's perspective.[15] Moreover, benchmarking must take into account the multimodal nature of traditional therapy (whereby human therapists rely on visual and vocal cues) and the multifaceted nature of therapy services (which include human therapists providing homework and doing case management). These are all currently understudied by existing LLM benchmark efforts and will require the close involvement of clinicians and those with lived experience.[16] It is also crucial to remember that benchmarks only measure model behavior and cannot be used to make claims about human clinical outcomes.

Comprehensive benchmarks will also require larger datasets sourced from real user conversations that are multiturn (i.e., involving multiple exchanges that reflect realistic use cases as opposed to single question-and-answer formats), demographically balanced (taking into account various personality types and neurodivergent subgroups), and representative of a range of mental illnesses. This will require finding more privacy-protecting ways to incorporate user data into benchmark development.

Recommendation 2: Require chatbot developers to provide end points for user-facing models that can be used for benchmarking. Most LLM benchmarking research is done using an application programming interface (API) end point. Developers use these to ensure their models' reliable and scalable performance while academic researchers rely on them to conduct systematic research on commercially available LLMs. Yet the API model differs from the version users see when they use a chatbot: User-facing models are updated frequently and often have customization features (including user-selectable "personas," memory, and user-editable prompts) that make it virtually impossible to predict model behavior. Benchmarking performance on API end points may thus have little relevance to the impact that actual users see when they interact with the user-facing model. Commercial providers often do not make available an end point for the user-facing version of their models, nor do they specify precisely how they differ from API models. Academic and governmental researchers and third-party evaluators must have access to these user-facing model end points in order to run the evaluations we call for above.

Based on LLM for High-Functioning Autistic Adolescent Psychological Counseling," *arXiv preprint*, November 12, 2023, https://arxiv.org/abs/2311.09243.

[15] Zhang, "CBT-Bench"; Andre Ye et al., "Language Models as Critical Thinking Tools: A Case Study of Philosophers," *arXiv preprint*, last revised August 7, 2024, https://arxiv.org/abs/2404.04516; Angelina Wang et al., "Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups," *arXiv preprint*, last revised February 3, 2025, https://arxiv.org/abs/2402.01908; Andrea Cuadra et al., "The Illusion of Empathy? Notes on Displays of Emotion in Human-Computer Interaction," *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, May 11, 2024: 1–18, https://dl.acm.org/doi/10.1145/3613904.3642336.

[16] Jina Suh et al., "Rethinking Technology Innovation for Mental Health: Framework for Multi-Sectoral Collaboration," *Nature Mental Health* 2, no. 5 (May 2024): 478–88, https://www.nature.com/articles/s44220-024-00232-2.

Recommendation 3: Institute reporting requirements for performance evaluations and safety protocols. While it is laudable that some companies have recently begun to release mental-health-relevant statistics, greater transparency is needed.[17] Such reporting must better identify internal evaluation methods, including what classifiers were used in testing and how clinical experts were consulted. Greater transparency regarding whether, when, and how LLM providers step in to de-escalate conversations or guide users to professional care and crisis hotlines is also needed. California's recently passed companion chatbot bill, SB243, which requires chatbot developers to annually report suicide protocols and related statistics to the state's Office of Suicide Prevention, is a step in the right direction.[18] However, such reporting must become more standardized and be made available to academic researchers to improve public understanding of the safety of these models.

Recommendation 4: Designate a trusted third-party evaluator for AI-mental health chatbots. Early efforts to create evaluation methods and benchmarks for AI chatbots used in mental health contexts (including the above-mentioned Spiral-Bench) have thus far been led by independent researchers. To our knowledge, there currently exists no public or nonprofit institution or infrastructure for conducting third-party evaluations of mental-health-related AI chatbot performance. The National Institute of Standards and Technology's (NIST) face recognition vendor test could be one model.[19] Facial recognition technology vendors submit their models to NIST, which runs tests to evaluate their performance on a range of characteristics such as accuracy and demographic bias. Designating a trusted third-party evaluator ensures impartiality and fairness, rather than simply relying on companies to self-report their metrics.

Advancing the design of mental health chatbots grounded in behavioral sciences

Every day, chatbot developers make a host of behavioral design choices — without professional guidance or regulatory oversight — that affect millions of users who turn to chatbots for mental health support.

One important feature of LLM-powered chatbots, compared to other AI-enabled medical devices, is that having conversations with a chatbot adds significant complexity to the user interaction. For instance, the latest version of ChatGPT touts a range of options to "customize ChatGPT"'s tone and style," from candid to friendly and professional.[20] Of course, these persona

[17] OpenAI, "Strengthening ChatGPT's responses in sensitive conversations," October 27, 2025, https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/.
[18] Colin Lecher, "New California Law Forces Chatbots to Protect Kids' Mental Health," *Cal Matters*, October 13, 2025, https://calmatters.org/economy/technology/2025/10/newsom-signs-chatbot-regulations/.
[19] NIST, Face Recognition Vendor Test (FRVT), https://www.nist.gov/programs-projects/face-projects.
[20] OpenAI, "GPT-5.1: A Smarter, More Conversational ChatGPT," November 12, 2025, https://openai.com/index/gpt-5-1/.

options are not new — many social chatbots over the past decade have been engineered to optimize for engagement with certain target user groups.[21] But what is new is how many users intentionally turn to such chatbots for mental health support. There are increasing reports of users developing parasocial relationships with these chatbots, such as feelings of emotional dependence that can lead to delusions, which have been cited in wrongful death lawsuits filed against OpenAI and Character.ai.[22] Model anthropomorphization is a related problem: Many concerning cases have been documented of models claiming to be human or having emotions, or even purporting to be a "flesh-and-blood [licensed] therapist."[23]

Chatbots that are designed to serve as AI therapists or counselors, or otherwise provide mental health support, should be subject to design constraints. Future regulatory mechanisms should be oriented around harm reduction, on the assumption that consumers will continue to seek out ways to use LLMs for mental health support, even if there are regulations and safeguards.

Recommendation 5: Mandate companies to designate products designed for therapeutic uses. Chatbot developers should be forced to differentiate their products. General-purpose LLMs are widely used for mental health purposes, but they are optimized for engagement at the expense of mental health (e.g., by allowing persona changes, offering role-play, and otherwise optimizing for prolonged user engagement). Forcing companies to designate a separate "therapy" version of their chatbots and identify carve-outs for use-cases that demand greater safety protocols would allow safer, more tailored design as well as more targeted government oversight. These designated products should then have a responsibility to optimize for safety in addition to sustained user engagement.

Recommendation 6: Ensure chatbot design that prevents AI sycophancy and parasocial relationships. Decades of research have shown that people naturally anthropomorphize technology.[24] More recently, researchers have documented LLMs' "sycophantic" characteristics, which present in behaviors that range from overly flattering comments to overtly encouraging delusions.[25] Resulting harmful symptoms among users are widely being referred to as "AI

[21] For example, Microsoft's XiaoIce was designed as an empathetic AI companion. *See* Li Zhou et al., "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot," *Computational Linguistics* 46, no. 1 (March 2020): 53–93, https://aclanthology.org/2020.cl-1.2/.
[22] Hill, "Lawsuits Blame ChatGPT for Suicides and Harmful Delusions."
[23] Andrew R. Chow and Angela Haupt, "A Psychiatrist Posed As a Teen with Therapy Chatbots. The Conversations Were Alarming," *Time*, June 12, 2025, https://time.com/7291048/ai-chatbot-therapy-kids/.
[24] Joseph Weizenbaum, "ELIZA: A Computer Program for the Study of Natural Language Communication between Man and Machine," *Communications of the ACM* 9, no. 1 (January 1966): 36–45, https://dl.acm.org/doi/10.1145/365153.365168; Clifford Nass et al., "Computers Are Social Actors," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, April 24, 1994: 72–78, https://dl.acm.org/doi/pdf/10.1145/191666.191703.
[25] Myra Cheng et al., "Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence," *arXiv preprint*, October 1, 2025, https://arxiv.org/abs/2510.01395; Myra Cheng et al., "ELEPHANT: Measuring and Understanding

psychosis."[26] While we still lack definitive evidence to design thoughtful mitigation measures, it is clear that developers must address these issues during the chatbot design phase. LLMs are sycophantic because they are trained to be agreeable, leading them to refrain from pushing back against the user and to entertain delusions or suicidal thoughts.

Mental health chatbots should be deliberately designed to maintain an appropriate, professional stance. Human mental healthcare professionals have codes of ethics designed to preempt therapists from forming relationships with their clients and prevent numerous types of harm. Chatbot developers, therefore, should not incorporate behavioral design features that might encourage client parasocial relationships and model anthropomorphization. We are not aware of specific technical work on best practices or guidelines for mental health chatbot features (e.g., an ideal therapist persona). Novel research is urgently needed to determine what constitutes appropriate chatbot engagement levels, what appropriate professional responses look like in the chatbot context, as well as what kinds of guardrails are feasible as tools to prevent parasocial relationships with chatbots.

***

Rigorously developed and scientifically validated AI tools have the potential to help close the mental healthcare gap in the United States. However, the present state of affairs — where such applications lack a proper regulatory framework and mandated guardrails — means that individuals who seek mental healthcare from AI are using products that have not been designed for mental healthcare and are not rigorously tested. We urge the FDA to consider implementing these recommendations swiftly but also with sufficient flexibility to allow for responding to the evolving AI landscape.

Sincerely,

Desmond C. Ong, PhD
Assistant Professor, Department of Psychology, The University of Texas at Austin
Affiliate, Texas Behavioral Science & Policy Institute, UT Austin

Jared Moore
PhD Student, Department of Computer Science, Stanford University

---

Social Sycophancy in LLMs," *arXiv preprint*, last revised September 29, 2025, https://arxiv.org/abs/2505.13995; Moore et al., "Expressing Stigma and Inappropriate Responses."

[26] More Perfect Union, "We Investigated AI Psychosis. What We Found Will Shock You," YouTube, accessed November 26, 2025, https://www.youtube.com/watch?v=zkGk_A4noxI; Huet and Metz, "The Chatbot Delusions."

Nicole Martinez-Martin, JD, PhD
Assistant Professor, Department of Pediatrics and, by courtesy, Department of Psychiatry and Behavioral Sciences, Stanford University
Faculty Affiliate, Stanford Institute for Human-Centered Artificial Intelligence

Caroline Meinhardt
Policy Research Manager, Stanford Institute for Human-Centered Artificial Intelligence

Eric Lin, MD
Addiction Psychiatrist, Medical Informatician, and AI Consultant

William Agnew, PhD
Carnegie Bosch Postdoctoral Fellow, Carnegie Mellon University