Is This Chatbot Safe and Evidence-Based? Establishing a Framework for the Critical Evaluation of Gen AI Mental Health Chatbots

Acacia Parks, Eoin Travers, Ramesh Perera-Delcourt, Max Major, Marcos Economides, Phil Mullan

Submitted to: Journal of Participatory Medicine
on: December 02, 2024

Table of Contents

Is This Chatbot Safe and Evidence-Based? Establishing a Framework for the Critical Evaluation of Gen AI Mental Health Chatbots

Acacia Parks[1] MBA, PhD; Eoin Travers[1] PhD; Ramesh Perera-Delcourt[1] PhD; Max Major[1] PhD; Marcos Economides[1] PhD; Phil Mullan[1] BSc

[1]Unmind London GB

Corresponding Author:
Acacia Parks MBA, PhD
Unmind
140 Borough High St
London
GB

Abstract

The proliferation of AI mental health chatbots, such as those on platforms like OpenAI's GPT Store and Character.AI, raises issues of safety, effectiveness, and ethical use; they also raise an opportunity for patients and consumers to ensure AI tools clearly communicate how they meet their needs. While many of these tools claim to offer therapeutic advice, their unregulated status and lack of systematic evaluation create risks for users, particularly vulnerable individuals. This viewpoint article highlights the urgent need for a standardized framework to assess and demonstrate the safety, ethics, and evidence basis of AI chatbots used in mental health contexts. Drawing on clinical expertise, research, co-design experience, and WHO guidance, the authors propose key evaluation criteria: adherence to ethical principles, evidence-based responses, conversational skills, safety protocols, and accessibility. Implementation challenges, including setting output criteria without one 'right answer', evaluating multi-turn conversations, and involving experts for oversight at scale, are explored. The authors advocate for greater consumer engagement in chatbot evaluation to ensure these tools address user needs effectively and responsibly, emphasizing the ethical obligation of developers to prioritize safety and a strong base in empirical evidence.

(JMIR Preprints 02/12/2024:69534)
DOI: https://doi.org/10.2196/preprints.69534

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Is This Chatbot Safe and Evidence-Based?
Establishing a Framework for the Critical Evaluation of Gen AI Mental Health Chatbots

Acacia C. Parks
Max Major
Marcos Economides
Ramesh Perera-Delcourt
Eoin Travers
Phil Mullan

Unmind Ltd, London, United Kingdom

For Submission to the Journal of Participatory Medicine
Consumer Use of Artificial Intelligence for Health Special Issue

Key Words: GenAI; mental health; chatbot; ethics; evals

ABSTRACT

The proliferation of AI mental health chatbots, such as those on platforms like OpenAI's GPT Store and Character.AI, raises issues of safety, effectiveness, and ethical use; they also raise an opportunity for patients and consumers to ensure AI tools clearly communicate how they meet their needs. While many of these tools claim to offer therapeutic advice, their unregulated status and lack of systematic evaluation create risks for users, particularly vulnerable individuals. This viewpoint article highlights the urgent need for a standardized framework to assess and demonstrate the safety, ethics, and evidence basis of AI chatbots used in mental health contexts. Drawing on clinical expertise, research, co-design experience, and WHO guidance, the authors propose key evaluation criteria: adherence to ethical principles, evidence-based responses, conversational skills, safety protocols, and accessibility. Implementation challenges, including setting output criteria without one 'right answer', evaluating multi-turn conversations, and involving experts for oversight at scale, are explored. The authors advocate for greater consumer engagement in chatbot evaluation to ensure these tools address user needs effectively and responsibly, emphasizing the ethical obligation of developers to prioritize safety and a strong base in empirical evidence.

Is This Chatbot Safe and Effective For Me To Use?
Establishing a Framework for the Critical Evaluation of Mental Health Chatbots

The internet is flooded with mental health resources, and one of the most common emerging formats is the AI (artificial intelligence) chatbot. A recent Forbes article examines the launch of OpenAI's GPT store, which allows users to post chatbots for ready use by others, and found many were intended for mental health advisory purposes; another 3 million or so general purpose chatbots are not intended specifically for mental health purposes, but would take on that role if prompted [1]. For example, a quick Google search for "Character.AI" and "therapist" yields a link to a Character.AI bot who says they have "been working in therapy since 1999… [are] a Licensed Clinical Professional Counselor (LCPC)... [and are] trained to provide EMDR treatment in addition to Cognitive Behavioral (CBT) therapies." A small disclaimer at the bottom states, "This is A.I. and not a real person. Treat everything it says as fiction." However, the boundary between reality and fiction can become quite blurry for consumers interacting with AI chatbots, as is illustrated by instances where deaths by suicide have been linked to chatbot usage [2].

This is particularly pertinent for chatbots which use Generative AI (GenAI). Although mental health chatbots have existed for some time, their increasing popularity is in part due to the rise of GenAI. In traditional chatbots, the user's interaction with the bot is typically governed by an explicitly programmed set of rules for choosing between pre-written responses. Generative chatbots, in contrast, are driven by powerful Large Language Models (LLMs), that produce customized responses to each user message, guided by the instructions written in the "system prompt" provided to the LLM. Generative chatbots provide much greater flexibility, at the cost of less predictable behavior.

The legality of such apps when used for mental health is questionable, as digital products that make medical claims - such as the ability to treat depression or anxiety - are considered in many countries to be medical devices. Medical devices are subject to requirements to show evidence of safety and effectiveness, as well as regulatory scrutiny. But the large majority of digital products that make these types of claims are not evaluated by regulatory bodies [3]. Somewhere in between "free for all" and "medical device" is a category of digital products that may provide advice responsibly without claiming they provide treatment. These chatbots can be considered 'general mental health support' bots, as opposed to conversational AI chatbots which have a specific purpose such as Triage [4]. Examples include Ada [5], Chai [6], Elomia [7], Mindspa [8], Nuna [9], Serenity [10], Stresscoach [11], Woebot [12], Wysa [13], and Youper [14,15], as well as newer entrants Ebb (Headspace [16]) and Nova (Unmind [17]). Because these and other similar chatbots don't rise to the level of a medical device, regulatory bodies (e.g. FDA) do not govern the claims made about what the chatbots do. Consumers are therefore left to navigate this landscape without guidance on what makes a chatbot safe and effective. However, there is currently no legal, academic or industry agreed standard or method for doing this in a way that enables consumers to be meaningful, active collaborators in their own care.

We argue that companies producing AI mental health products intended for general use should demonstrate, in some systematic and objective way, that the products they provide to consumers are safe and deliver advice that is evidence-based. We argue that doing so is an ethical obligation to consumers, as well as something (quite rightly) expected of digital mental health interventions by both users and providers who recommend digital products. To empower consumers and the public to accurately assess the risks and benefits of using AI for self-care, there needs to be a clear, accessible framework for evidencing how the chatbot addresses the needs and concerns of the individual user. This framework will also need to be meaningful and acceptable to potential

gatekeepers of access to AI, such as therapist referrers or employer health benefit providers.

What Criteria Should Generative, General Mental Health Chatbots Be Evaluated On?

Evaluating mental health-related chatbots is a particular challenge, due to the sensitive nature of mental health, and the consequences of providing poor quality responses to potentially vulnerable users discussing sensitive topics. Based on our shared experience in clinical practice, mental health research codesign/participatory involvement in research and building AI-powered products, and on the World Health Organisation's (WHO) guidance on 'Ethics & Governance of Artificial Intelligence for Health (2024)' [18], we propose that mental health AI chatbots should adhere to a version of the criteria below.

Table 1. Criteria for evaluating performance of an AI mental health chatbot

Criteria	**Definition**
Be Ethical	Responses should benefit users while avoiding harm, be just and fair, promote user autonomy, and allow for transparent, informed understanding of their basis
Be Safe	Clear rules governing a chatbot's behaviour when there is risk of physical or psychological harm to the user or to others must be set and adhered to. These should establish the chatbot's remit, including signposting to external resources and not providing medical diagnosis or treatment or producing any outputs that would constitute use as a regulated medical device.
Be Accessible	The chatbot should be accessible to the user, including support for the user's native language where possible and appropriate accommodation for the user's verbal comprehension skills.
Follow the Evidence-Base	Responses should be grounded in the established scientific literature.
Apply Core Coaching Skills	The chatbot should display strong conversational skills, and apply conversational techniques including goal identification, alliance building, and empathetic inquiry.

Whatever criteria we use and whatever thresholds we set for expected performance of a chatbot, they should have real-world impact and reflect what matters most to users, including: perceived relevance and usefulness, privacy and confidentiality [19], and human therapist personal attributes valued by consumers that may be replicable by AI chatbots, such as being respectful, confident, warm, and interested [20,21].

How Could A Benchmark Be Implemented?

With the explosion in applications of GenAI, there is greater emphasis placed on "evals": systematic approaches to evaluating whether the outputs of the AI system are appropriate for the task at hand before they are rolled out to users [22,23]. Evals will typically consist of a collection of test

inputs to the AI system, and criteria or scoring rules by which to evaluate the outputs. There are some scenarios where the accuracy of outputs may be evaluated directly, for instance by comparing against a predefined target or using pattern matching. In other cases, for instance in applications involving classification, data retrieval, or summarisation, outputs can be compared against targets using statistical metrics such as precision and recall.

However, in many applications of GenAI, particularly those involving chatbots, there is no meaningful "right answer" for the chatbot to give. In these cases, we must instead evaluate outputs against a rubric or set of qualitative criteria. Criteria might include formatting features (e.g., uses markdown), linguistic style (e.g., level of formality), tone of voice (e.g., level of warmth) or more abstract features (e.g., shows empathy). This approach is used in the reinforcement learning phase of training modern AI LLMs, where models will generate multiple candidate responses to a given question, the preferred response is identified using predefined criteria, and this feedback is used to adjust the model to make such a response more likely [24,25], but is equally useful in evaluating models after training.

Evaluations against criteria can be performed either by human annotators or by additional AI systems. Expert human annotators can bring deep clinical expertise and nuanced understanding to their evaluations [25,26]. However, this approach is extremely resource intensive, and may suffer from unreliability or inconsistency, particularly when annotating large data sets [27]. An emerging alternative is the "LLM-as-a-judge" approach [28,29], where these evaluations are performed by an LLM. To work reliably, this approach requires an additional process of comparing LLM-generated evaluations against high-quality human evaluations, and modifying the instruction prompt used by the LLM to align and calibrate the human and AI judgements.

Writing criteria against which to evaluate AI-generated responses is a deceptively difficult task, requiring a deep understanding of the domain, and the likely behaviours of both the users and the chatbot. It is increasingly recognised that the implicit criteria used by human annotators evolve as they are exposed to a greater variety of data [29]. It is considered best practice [29] to write these criteria iteratively, with expert judges continuously reviewing real user data alongside the previous generation of LLM-judged evals in order to produce criteria that better define how a chatbot should behave.

For chatbots, evals based on single interactions (a message and a response) may fail to capture important dynamics that emerge over multiple turns in a conversation. A promising approach is to use an additional AI system to play the role of the user interacting with the target chatbot in order to simulate multi-turn "bot-to-bot" conversations. This approach has its challenges. If we intend to generalise from the chatbot's responses in these simulated conversations to how the chatbot would respond in real interactions with humans, we must ensure that the messages from the simulated user are representative of the range of messages that would be sent by real users. Multi-turn conversations can also go down many more diverging paths than single interactions, and so a large number of simulated conversations under the same conditions may be needed to allow for the variance in outcomes.

The Role of the Consumer

Much research to date has focused on using professional experts, not healthcare users, to evaluate chatbots. Although inconsistent, research has shown that coproduction of digital mental health interventions can improve their utility [30]. Similarly to how there is a need for guidelines around user involvement in intervention development [31], we believe that the implementation of the above framework would benefit from healthcare consumers not just contributing to the evaluation

criteria, but being involved in rating chatbot conversations to calibrate the automated testing systems. This would ensure that health chatbots are evaluated in line with not just what previous research has demonstrated is important to consumers, but also what is currently most relevant given this technology is emergent.

Conclusion

Digital mental health is rife with products that are unhelpful at best, and compromise consumer safety at worst. In order to realize the potential of AI for mental health, it is recognised that all stakeholders need to be involved in its development and regulation [32]. We have argued for the importance of evaluating AI mental health chatbots, even in a non-regulated context, objectively, with a common set of criteria that can provide guidance for consumers and practitioners on which products are safe and evidence based. We provide a framework to start, and highlight some of the key challenges to implementing such a framework. By involving consumers in the evaluation process, and addressing their needs during development, the true promise of AI can be realized for all healthcare users.

Author Roles

ACP wrote the first draft and oversaw the integration of input from the remaining authors. MM, ME, RPD, ET, and PM contributed substantively to the main text and provided input on clarity and concision throughout. ME prepared the manuscript for publication.

Acknowledgments

Amanda Woodward provided instrumental support in collecting and organizing citations.

Conflicts of Interest

All authors were employed by Unmind Ltd at the time this viewpoint was written, and MM, ME, RPD, ET, and PM own share options at Unmind Ltd. Unmind Ltd is the creator of Nova, one of the GenAI chatbot products discussed in this article.

Abbreviations

AI: Artificial Intelligence
GenAI: Generative Artificial Intelligence
LLM: Large Language Model
WHO: World Health Organization

References

1. Eliot, L. Newly Launched GPT Store Warily Has ChatGPT-Powered Mental Health AI Chatbots That Range From Mindfully Serious To Disconcertingly Wacko. 2024. Available from: https://www.forbes.com/sites/lanceeliot/2024/01/14/newly-launched-gpt-store-warily-has-chatgpt-powered-mental-health-ai-chatbots-that-range-from-mindfully-serious-to-disconcertingly-wacko/
2. Fraser, H. Deaths linked to chatbots show we must urgently revisit what counts as ‘high-risk’ AI. 2024. Available from: https://theconversation.com/deaths-linked-to-chatbots-show-we-must-urgently-revisit-what-counts-as-high-risk-ai-242289
3. Freyer O, Wrona KJ, De Snoeck Q, Hofmann M, Melvin T, Stratton-Powell A, Wicks P, Parks AC, Gilbert S. The regulatory status of health apps that employ gamification. Sci Rep 2024 Sep

9;14(1):21016. doi: 10.1038/s41598-024-71808-2
4. Rollwage M, Habicht J, Juchems K, Carrington B, Hauser TU, Harper R. Conversational AI facilitates mental health assessments and is associated with improved recovery rates. BMJ Innov 2024 Jan;10(1–2):4–12. doi: 10.1136/bmjinnov-2023-001110
5. Ada. Available from: https://ada.com/
6. Chai. Available from: https://www.chai-research.com/
7. Elomia Health. Available from: https://elomia.com/
8. Mindspa. Available from: https://mindspa.me/en/
9. Nuna. Available from: https://nuna.ai/
10. Serenity. Available from: https://www.serenityfeels.co.in/
11. Stresscoach. Available from: https://www.stresscoach.app/
12. Woebot Health. Available from: https://woebothealth.com/
13. Wysa. Available from: https://www.wysa.com/
14. Youper. Available from: https://www.youper.ai/
15. Haque MDR, Rubya S. An Overview of Chatbot-Based Mobile Mental Health Apps: Insights From App Description and User Reviews. JMIR Mhealth Uhealth 2023 May 22;11:e44838. doi: 10.2196/44838
16. Ebb by Headspace. Available from: https://www.headspace.com/ai-mental-health-companion
17. Nova by Unmind. Available from: https://unmind.com/feature-nova
18. Geneva: World Health Organization. Ethics and governance of artificial intelligence for health. Guidance on large multi-modal models. 2024 p. Licence: CC BY-NC-SA 3.0 IGO.
19. Borghouts J, Eikey E, Mark G, De Leon C, Schueller SM, Schneider M, Stadnick N, Zheng K, Mukamel D, Sorkin DH. Barriers to and Facilitators of User Engagement With Digital Mental Health Interventions: Systematic Review. J Med Internet Res 2021 Mar 24;23(3):e24387. doi: 10.2196/24387
20. Ackerman SJ, Hilsenroth MJ. A review of therapist characteristics and techniques positively impacting the therapeutic alliance. Clinical Psychology Review 2003 Feb;23(1):1–33. doi: 10.1016/S0272-7358(02)00146-0
21. Naher J. Can ChatGPT provide a better support: a comparative analysis of ChatGPT and dataset responses in mental health dialogues. Curr Psychol 2024 Jul;43(28):23837–23845. doi: 10.1007/s12144-024-06140-z
22. Ganguli D, Schiefer N, Favaro M, Clark J. Challenges in evaluating AI systems. 2023. Available from: https://www.anthropic.com/index/evaluating-ai-systems
23. Yan, E., Bischof, B., Frye, C., Husain, H., Liu, J., Shankar, S. Applied LLMs - What We've Learned From A Year of Building with LLMs. Applied LLMs 2024 Jun; Available from: https://applied-llms.org/
24. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, Chen A, Goldie A, Mirhoseini A, McKinnon C, Chen C, Olsson C, Olah C, Hernandez D, Drain D, Ganguli D, Li D, Tran-Johnson E, Perez E, Kerr J, Mueller J, Ladish J, Landau J, Ndousse K, Lukosuite K, Lovitt L, Sellitto M, Elhage N, Schiefer N, Mercado N, DasSarma N, Lasenby R, Larson R, Ringer S, Johnston S, Kravec S, Showk SE, Fort S, Lanham T, Telleen-Lawton T, Conerly T, Henighan T, Hume T, Bowman SR, Hatfield-Dodds Z, Mann B, Amodei D, Joseph N, McCandlish S, Brown T, Kaplan J. Constitutional AI: Harmlessness from AI Feedback. arXiv; 2022. doi: 10.48550/ARXIV.2212.08073
25. Ziegler DM, Stiennon N, Wu J, Brown TB, Radford A, Amodei D, Christiano P, Irving G. Fine-Tuning Language Models from Human Preferences. arXiv; 2019. doi: 10.48550/ARXIV.1909.08593
26. Qiu H, Li A, Ma L, Lan Z. PsyChat: A Client-Centric Dialogue System for Mental Health Support. 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD) Tianjin, China: IEEE; 2024. p. 2979–2984. doi:

10.1109/CSCWD61410.2024.10580641
27. Sylolypavan A, Sleeman D, Wu H, Sim M. The impact of inconsistent human annotations on AI driven clinical decision making. npj Digit Med 2023 Feb 21;6(1):26. doi: 10.1038/s41746-023-00773-3
28. Yan Z. Evaluating the Effectiveness of LLM-Evaluators (aka LLM-as-Judge). eugeneyan.com 2024 Aug; Available from: https://eugeneyan.com/writing/llm-evaluators/
29. Shankar S, Zamfirescu-Pereira JD, Hartmann B, Parameswaran AG, Arawjo I. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. arXiv; 2024. doi: 10.48550/ARXIV.2404.12272
30. Brotherdale R, Berry K, Branitsky A, Bucci S. Co-producing digital mental health interventions: A systematic review. DIGITAL HEALTH 2024 Jan;10:20552076241239172. doi: 10.1177/20552076241239172
31. Bernaerts S, Van Daele T, Carlsen CK, Nielsen SL, Schaap J, Roke Y. User involvement in digital mental health: approaches, potential and the need for guidelines. Front Digit Health 2024 Aug 22;6:1440660. doi: 10.3389/fdgth.2024.1440660
32. Hopkin G, Branson R, Campbell P, Coole H, Cooper S, Edelmann F, Gatera G, Morgan J, Salmon M. Building robust, proportionate, and timely approaches to regulation and evaluation of digital mental health technologies. The Lancet Digital Health 2024 Nov;S2589750024002152. doi: 10.1016/S2589-7500(24)00215-2