Standardizing and Scaffolding Healthcare AI-Chatbot Evaluation

Yining Hua,[1,2] Winna Xia,[2] David W. Bates,[3†] Luke Hartstein,[4†] Hyungjin Tom Kim,[5†] Michael Lingzhi Li,[6†] Benjamin W. Nelson, PhD,[2,7,8†] Charles Stromeyer IV,[9†] Darlene King,[10] Jina Suh,[11†] Li Zhou,[3†] John Torous[2,7*]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2]Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, USA
[3]Division of Internal Medicine, Brigham and Women's Hospital, Boston, MA, USA
[4]Department of Psychiatry, Thomas Jefferson University, Philadelphia, PA, USA
[5]Department of Psychiatry and Human Behavior, Alpert Medical School of Brown University, Providence, RI, USA
[6]Technology and Operations Management, Harvard Business School
[7]Department of Psychiatry, Harvard Medical School, Boston, MA, USA
[8]Verily Life Sciences, San Francisco, CA, USA
[9]Patient Advisory Board, Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, USA
[10]Department of Psychiatry, University of Texas Southwestern, Houston, TX, USA
[11]Microsoft Research, Redmond, WA

[†]*Equal contribution, ordered alphabetically*
[*]*Correspondence: jtorous@bidmc.harvard.edu*

Abstract

The rapid rise of healthcare chatbots, valued at $787.1 million in 2022 and projected to grow at 23.9% annually through 2030, underscores the need for robust evaluation frameworks. Despite their potential, the absence of standardized evaluation criteria and rapid AI advancements complicate assessments. This study addresses these challenges by developing a the first comprehensive evaluation framework inspired by health app regulations and integrating insights from diverse stakeholders. Following PRISMA guidelines, we reviewed 11 existing frameworks, refining 271 questions into a structured framework encompassing three priority constructs, 18 second-level constructs, and 60 third-level constructs. Our framework emphasizes safety, privacy, trustworthiness, and usefulness, aligning with recent concerns about AI in healthcare. This adaptable framework aims to serve as the initial step in facilitating the responsible integration of chatbots into healthcare settings.

Introduction

The rapid rise of chatbots, also known as conversational agents, has garnered substantial interest in the healthcare market. Valued at $787.1 million in 2022, the global healthcare chatbot market is expected to grow at an annual rate of 23.9% from 2023 to 2030.[1] This expansion is driven by the increasing demand for virtual health assistance, growing collaborations between healthcare providers and industry players, and the acceleration prompted by the COVID-19 pandemic. For example, over 1,000 healthcare organizations worldwide developed COVID-19-specific chatbots using Microsoft's Healthcare Bot service to manage patient inquiries and reduce the burden on medical staff.[2] Entering the age of generative artificial intelligence (AI), healthcare chatbots have received even more attention since they enable human-level fluent conversations, have reached physician-level performance on board residency examinations[3] and comparable performance on other medical examinations and questions[4-,6] and offer easy ways to train and adapt.

But despite their popularity and potential, evaluating healthcare chatbots poses many challenges.[7-9] A lack of standardized evaluation approaches has led to diverse and inconsistent methods, making comparing chatbot performance difficult. Rapid technological advancements, particularly in generative AI, outpace existing regulatory frameworks[10], complicating the establishment of evaluation standards. These new chatbots utilizing generative AI are not constrained by decision trees and are often built on top of larger models, meaning both the output and foundation are not stable. With such a moving target for evaluation, there is no widely accepted guideline or framework for evaluating healthcare chatbots. Developers lack a guide for assessment,[11] and users often rely on company advertisements or marketing claims.

Several evaluation frameworks[12-22] have emerged in response to these challenges over the last few years, particularly following the popularity of generative AI. These frameworks vary: some review existing works and regroup metrics into a new structure, others adapt non-healthcare evaluation frameworks for this field, and some

focus on narrow sub-directions such as specific specialties or chatbot types. Given the need for a general guiding evaluation framework, a novel approach is necessary. Inspired by a framework[23] for evaluating health apps, which has now been adopted by the American Psychiatric Association (APA), we crafted a general evaluation framework integrating a literature review and broad stakeholder analyses. This approach involves the perspectives of developers, clinicians, patients, and policymakers to create a comprehensive evaluation structure.

Methods

As healthcare chatbots face a variety of users, there is no single way to evaluate a chatbot. Factors such as safety and privacy, user preferences, technology literacy, accessibility, and treatment goals are crucial in determining the most suitable evaluation method. In addressing these issues, organizations like the Coalition for Health AI (CHAI) have been working on designing guidelines for trustworthy AI. In April 2023, a group of experts representing diverse stakeholders crafted a blueprint for trustworthy AI implementation guidance.[24] This blueprint includes seven aspects of trustworthy AI in healthcare: usefulness, safety, accountability and transparency, explainability and interpretability, fairness, security and resilience, and enhanced privacy. But this framework serves more as a theoretical foundation rather than an empirical evaluation framework, and its similarity or overlap with other frameworks remains unclear. Building on the construct definitions in this blueprint and existing evaluation frameworks, we 1) identified a total of 11 evaluation frameworks, 2) extracted all individual questions from these frameworks, 3) removed redundant and non-relevant questions, 4) mapping the remaining questions to CHAI constructs, their subcategories, and constructs not covered by CHAI's blueprint, 5) improved the evaluation framework structure with stakeholders, including healthcare providers, patients, technology developers, epidemiologists, and policymakers, and 6) further merged and rephrased questions based on assigned constructs.

Due to the absence of a comprehensive review of healthcare chatbot evaluation frameworks, we followed the PRISMA guidelines for selecting and reviewing papers (Appendix A) and gathered 356 questions from the 11 evaluation frameworks (Appendix B). After removing redundant and non-relevant questions (n=35, process detailed in Appendix C), the remaining questions were analyzed for face and construct validity and mapped onto seven priority levels, reflecting the CHAI framework. Subcategories were identified by further clustering questions and reorganizing the framework structure, merging and dividing overlapping questions. This process was modeled as a qualitative factor analysis, where all authors examined and reached a consensus on how the questions were categorized. Based on this refined constructs and framework structure, questions were re-analyzed to form a final list (n=271, listed in Appendix D).

Results

The final framework (first two levels shown in figure 1; full framework shown in Appendix E) represents three priority level constructs, 18 second-level constructs, and 60 third-level constructs. The 271 questions covered 56 third-level constructs. Among these questions, Design and Operational Effectiveness accounted for 108 (40%) questions. Trustworthiness and Usefulness accounted for a similar weight of 107 questions each (39%). The most fundamental level of Safety, Privacy, and Fairness included 56 questions (21%). Subcategories have different levels of granularity, with some categories having only one question and others having many (Appendix F).

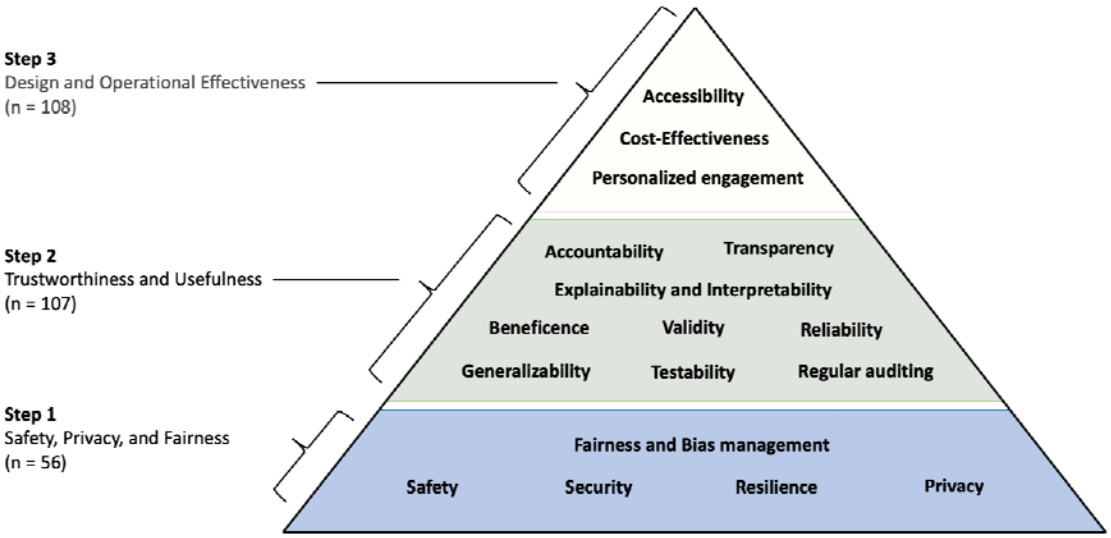

Figure 1: Pyramid for healthcare chatbot evaluation framework. Priority-level constructs are displayed on the left, with second-level constructs within the pyramid.

The rise of generative AI, such as ChatGPT, has expanded interest in healthcare chatbots, placing a pressing need for robust evaluation guidance. Yet the emergence of so many frameworks may create more uncertainty. By assessing the details of numerous frameworks, we were able to simplify and unify different approaches to help inform decision-making. The current framework is designed to be flexible and serve different decision makers around different questions ranging from a designer seeking to create a new chatbot to a patient selecting one from the marketplace. Depending on the user and use case, a different weighting to each construct will be necessary in the same manner that ethical principles offer a scaffold to guide diverse decision making. Our analysis (see Appendix F) suggests that while most frameworks emphasize factors like user experience and task efficiency, stakeholder feedback suggests that a focus on safety and usefulness (see Figure 1) may better match user needs and concerns.

The pyramid structure, similar to Maslow's Hierarchy of Needs, serves as a visual reminder that evaluation may begin at the base, and progression is likely unnecessary if any level fails to meet the required standards. Still, the user may opt to approach the constructs and questions in any manner that suits their needs.The process of going through these questions will likely facilitate productive dialogue and reveal tensions that must be addressed by the user in order to make the optimal selection. Thus this structure does not itself perform an evaluation but rather serves as a scaffold for evaluation. The same chatbot will be evaluated differently depending on the user and their intent for use, reflecting the flexible nature of this framing. The detailed questions, summarized in Appendix E, are designed to encourage and facilitate dialogue among stakeholders, with responses contextualized within each stakeholder's unique situation. For instance, some chatbots may collect user conversation histories for training purposes by default. Some patients may find this unacceptable, while others may be comfortable with it. Similarly, developers focused on improving chatbot validity and reliability should not be compelled to conduct user feedback field studies if their research scope explicitly excludes user experience.

Discussion

Chatbots are increasingly widely used in healthcare, but no comprehensive framework for evaluating their performance has been available. We surveyed the existing frameworks and developed a new framework, using PRISMA guidelines, which we hope will enable future comparisons. This framework is designed to meet the myriad users, use cases, and advances around health AI chatbots by providing a flexible scaffolding to support informed decision making.

Our framework's foundation in safety, privacy, and fairness is well aligned with recent research raising concerns about these aspects of chatbots. A 2024 review of AI apps concluded these apps may cause harm associated with bias[25] and the 2023 real-world case of an AI chatbot for eating disorders giving dangerous information to users[26]

highlight the importance of Step 1 (see figure 1) in our framework. Not all AI chatbots are patient facing and the framework is relevant to scaffolding conversations about clinical documentation chatbots, differential diagnosis chatbots, even scheduling chatbots given the core aspects of the framework are relevant. For example, while efforts are underway to identify and address bias in conversational agents,[27] checking for and identifying bias in any chatbot is a productive first step in considering any conversational agent is a foundational step for avoiding harm.

Likewise, our framework's second step, trustworthiness, and usefulness, is grounded in recent research. From concerning trends of conversational agents drawing schizophrenia in a stigmatizing manner[28] to some chatbots providing details on self-harm and how to die by suicide,[29] it is critical to assess the trustworthiness and usefulness of conversational agents. Given most conversational agents today are trained on social media, not health data,[30] there is justified concern about the utility of information provided. Additionally, subtle errors can be mixed with correct responses that are difficult for even experts to detect[31]. While there are many approaches to determine trustworthiness and usefulness, and our framework does not dictate which should be employed, the structure ensures a focus on this critical issue.

Our framework also celebrates the success of conversational agents with step three considering factors like their often high degree of accessibility and efforts to personalize content. In placing step three after the prior two, our framework reminds the user to first consider the potential risks and appropriateness of the conversational agent. The majority of frameworks we assessed (see Appendix F) focused on the questions included here in step three. Our approach provides a complimentary means to consider these same questions but in the broader context of steps one and two.

Our framework offers several advantages by synthesizing insights from previous efforts into a new, synergistic model applicable across diverse health conditions and stakeholder groups. Unlike traditional methods that report isolated metrics, our framework reevaluates existing frameworks to distill and integrate them into a comprehensive general guiding framework. It is not designed to challenge or replace any framework and is flexible enough to incorporate new ones that will likely be developed.

A distinctive feature of our framework is its multi-level tree structure, mapping questions into granular constructs without assigning scores to individual questions. This approach facilitates future development of more detailed, domain-specific evaluation methods, using our framework as a reference or guide. Additionally, we aimed to maintain a consistent level of granularity across all levels of the framework, ensuring that each aspect of evaluation is addressed with equal thoroughness.

This approach has several limitations. The framework should be validated prospectively in different contexts to ensure that it is comprehensive and captures important dimensions. There may be additional dimensions that need to be added as the underlying technology quickly evolves, uncovering new issues.

Given the absence of a universal standard for evaluating healthcare chatbots, many parallel review tools have emerged, often failing to capture the full range of important considerations. Our framework addresses this gap, offering a comprehensive, adaptable tool for the evaluation of healthcare chatbots, which we hope will lead to responsible integration of chatbots into healthcare settings. Furthermore, we hope that this review could help guide policymakers to design effective evaluation regulation for healthcare chatbots, both to safeguard the quality of information and provide a clear roadmap for businesses worldwide to further develop tools that improve the quality, efficiency, and effectiveness of care.

This framework presents a starting point that will evolve. Next steps include fully exploring the needs of different users of health AI chatbots and their most common intent/goals. Exploring chatbots beyond the classical medical domains (e.g. nephrology, radiology) and understanding functions across the healthcare ecosystems from scheduling to crisis support will help ensure the framework is responsive to real world needs. Further work to expand the granularity of individual questions and their focus for users (e.g. developers vs clinicians) will help improve usability. Future endeavors will include a Delphi consensus based on these results in order to engage more stakeholders. Through these efforts, we hope to establish a more rigorous, inclusive, and widely adopted evaluation framework for healthcare chatbots, and enable "apples to apples" comparisons between them.

Funding

This study did not receive any funding.

Conflict of Interest

JT reports grants from Otsuka and is an advisor to Precision Mental Wellness, outside of the submitted work. DWB reports grants and personal fees from EarlySense, personal fees from CDI Negev, equity from ValeraHealth, equity from Clew, equity from MDClone, personal fees and equity from AESOP, personal fees and equity from Feelbetter, equity from Guided Clinical Solutions, and grants from IBM Watson Health, outside the submitted work. He has a patent pending (PHC-028564 US PCT), on intraoperative clinical decision support. BWN reports employment and equity ownership in Verily Life Sciences. JS is employed by Microsoft Research.
All other authors declare no competing interests.

References

1 Grand View Research. Healthcare Chatbots Market Size, Share & Trends Analysis Report By Component (Software, Services), By Application (Appointment Scheduling, Symptom Checking), By Deployment, By End-user, And Segment Forecasts, 2023 - 2030. San Francisco, CA: Grand View Research, Inc., 2024.

2 Bach D. How international health care organizations are using bots to help fight COVID-19. *Microsoft* 2020; published online April. https://news.microsoft.com/transform/how-international-health-care-organizations-are-using-bots-to-help-fight-covid-19/.

3 Katz U, Cohen E, Shachar E, *et al.* GPT versus Resident Physicians — A Benchmark Based on Official Board Scores. *NEJM AI* 2024; **1**: AIdbp2300192.

4 Meaney C, Huang RS, Lu K, Fischer AW, Leung FH, Kulasegaram K, Tzanetos K, Punnett A. Comparing the performance of ChatGPT and GPT-4 versus a cohort of medical students on an official University of Toronto undergraduate medical education progress test. medRxiv. 2023 Sep 14:2023-09.

5 Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS digital health. 2023 Feb 9;2(2):e0000198.

6 Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S, Payne P. Large language models encode clinical knowledge. Nature. 2023 Aug;620(7972):172-80.

7. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, Fries J, Shah NH. The shaky foundations of large language models and foundation models for electronic health records. npj Digital Medicine. 2023 Jul 29;6(1):135.

8 Torous J, Blease C. Generative artificial intelligence in mental health care: potential benefits and current challenges. *World Psychiatry* 2024; **23**: 1–2.

9 Organization WH. Ethics and governance of artificial intelligence for health: large multi-modal models. WHO guidance. World Health Organization, 2024.

10 Blumenthal D, Patel B. The Regulation of Clinical Artificial Intelligence. NEJM AI. 2024 Jul 12:AIpc2400545.

11. Zhou H, Liu F, Gu B, *et al.* A Survey of Large Language Models in Medicine: Progress, Application, and Challenge. 2024; published online July 10. DOI:10.48550/arXiv.2311.05112.

12 Denecke K, Warren J. How to Evaluate Health Applications with Conversational User Interface? *Stud Health Technol Inform* 2020; **270**: 976–80.

13 Denecke K. Framework for Guiding the Development of High-Quality Conversational Agents in Healthcare. *Healthcare* 2023; **11**: 1061.

14 Denecke K, May R. Developing a Technical-Oriented Taxonomy to Define Archetypes of Conversational Agents in Health Care: Literature Review and Cluster Analysis. *J Med Internet Res* 2023; **25**: e41583.

15 Liu C, Zowghi D, Peng G, Kong S. Information quality of conversational agents in healthcare. *Inf Dev* 2023; : 02666669231172434.

16 Martinengo L, Lin X, Jabir AI, *et al.* Conversational Agents in Health Care: Expert Interviews to Inform the Definition, Classification, and Conceptual Framework. *J Med Internet Res* 2023; **25**: e50767.

17 Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D'Alfonso S. To chat or bot to chat: Ethical issues with using chatbots in mental health. *Digit Health* 2023; **9**: 20552076231183542.

18 Xue J, Zhang B, Zhao Y, *et al.* Evaluation of the Current State of Chatbots for Digital Health: Scoping Review. *J Med Internet Res* 2023; **25**: e47217.

19 Abbasian M, Khatibi E, Azimi I, *et al.* Foundation metrics for evaluating effectiveness of healthcare

conversations powered by generative AI. *Npj Digit Med* 2024; **7**: 1–14.

20 Shlobin NA, Ward M, Shah HA, *et al.* Ethical Incorporation of Artificial Intelligence into Neurosurgery: A Generative Pretrained Transformer Chatbot-Based, Human-Modified Approach. *World Neurosurg* 2024; **187**: e769–91.

21 Ding H, Simmich J, Vaezipour A, Andrews N, Russell T. Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review. *J Am Med Inform Assoc* 2024; **31**: 746–61.

22 Nadarzynski T, Knights N, Husbands D, *et al.* Achieving health equity through conversational AI: A roadmap for design and implementation of inclusive chatbots in healthcare. *PLOS Digit Health* 2024; **3**: e0000492.

23 Henson P, David G, Albright K, Torous J. Deriving a practical framework for the evaluation of health apps. *Lancet Digit Health* 2019; **1**: e52–4.

24 Coalition for Health AI. Blueprint for Trustworthy AI: Implementation Guidance and Assurance for Healthcare. 2023; published online April. https://www.coalitionforhealthai.org/.

25 Wongvibulsin S, Yan MJ, Pahalyants V, Murphy W, Daneshjou R, Rotemberg V. Current State of Dermatology Mobile Applications With Artificial Intelligence Features. *JAMA Dermatol* 2024; **160**: 646–50.

26 Sharp G, Torous J, West ML. Ethical Challenges in AI Approaches to Eating Disorders. *J Med Internet Res* 2023; **25**: e50696.

27 Flores L, Kim S, Young SD. Addressing bias in artificial intelligence for public health surveillance. *J Med Ethics* 2024; **50**: 190–4.

28 King M. Harmful biases in artificial intelligence. *Lancet Psychiatry* 2022; **9**: e48.

29 De Freitas J, Cohen IG. The health risks of generative AI-based wellness apps. *Nat Med* 2024; **30**: 1269–75.

30 Hua Y, Liu F, Yang K, *et al.* Large Language Models in Mental Health Care: a Scoping Review. 2024; published online Jan 1. DOI:10.48550/arXiv.2401.02984.

31 Chen S, Kann BH, Foote MB, *et al.* Use of Artificial Intelligence Chatbots for Cancer Treatment Information. *JAMA Oncol* 2023; **9**: 1459–62.