

# 1 **Standardizing and Scaffolding Healthcare AI-Chatbot Evaluation**

2 Yining Hua, MSc<sup>1,2</sup>, Winna Xia, BS, BA<sup>2</sup>, David W. Bates, MD, MSc<sup>3†</sup>, George Luke Hartstein, MD, MBA<sup>4†</sup>,  
3 Hyungjin Tom Kim, MD<sup>5†</sup>, Michael Lingzhi Li, PhD<sup>6†</sup>, Benjamin W. Nelson, PhD,<sup>2,7,8†</sup> Charles Stromeier IV,<sup>9†</sup>  
4 Darlene King, MD<sup>10†</sup>, Jina Suh, PhD<sup>11†</sup>, Li Zhou, MD, PhD<sup>3†</sup>, John Torous, MD, MBI<sup>2,7\*</sup>

5 <sup>1</sup>Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

6 <sup>2</sup>Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, USA

7 <sup>3</sup>Division of Internal Medicine, Brigham and Women's Hospital, Boston, MA, USA

8 <sup>4</sup>Department of Psychiatry, Thomas Jefferson University, Philadelphia, PA, USA

9 <sup>5</sup>Department of Psychiatry and Human Behavior, Alpert Medical School of Brown University, Providence, RI,  
10 USA

11 <sup>6</sup>Technology and Operations Management, Harvard Business School

12 <sup>7</sup>Department of Psychiatry, Harvard Medical School, Boston, MA, USA

13 <sup>8</sup>Verily Life Sciences, San Francisco, CA, USA

14 <sup>9</sup>Patient Advisory Board, Department of Psychiatry, Beth Israel Deaconess Medical Center, Boston, MA, USA

15 <sup>10</sup>Department of Psychiatry, University of Texas Southwestern, Houston, TX, USA

16 <sup>11</sup>Microsoft Research, Redmond, WA

17 <sup>†</sup>*Equal contribution, ordered alphabetically*

18 <sup>\*</sup>*Correspondence:*

19 John Torous, MD, MBI

20 Department of Psychiatry

21 Beth Israel Deaconess Medical Center

22 Harvard Medical School

23 330 Brookline Ave

24 Boston, MA, 02446

25 United States

26 Phone: +1 (617) 6676-700

27 Email: [jtorous@gmail.com](mailto:jtorous@gmail.com)

28

29 # Abstract word count: 128

30 # Manuscript word count: 2,068

31 # Figures: 1

32 # Tables: 0

33 # References: 31

## 34 **Abstract**

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

## 44 **Introduction**

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

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

63 Several evaluation frameworks<sup>12-22</sup> have emerged in response to these challenges over the last few years,  
64 particularly following the popularity of generative AI. These frameworks vary: some review existing works and  
65 regroup metrics into a new structure, others adapt non-healthcare evaluation frameworks for this field, and some  
66 focus on narrow sub-directions such as specific specialties or chatbot types. Given the need for a general  
67 guiding evaluation framework, a novel approach is necessary. Inspired by a framework<sup>23</sup> for evaluating health  
68 apps, which has now been adopted by the American Psychiatric Association (APA), we crafted a general  
69 evaluation framework integrating a literature review and broad stakeholder analyses. This approach involves the  
70 perspectives of developers, clinicians, patients, and policymakers to create a comprehensive evaluation structure.

## 71 **Methods**

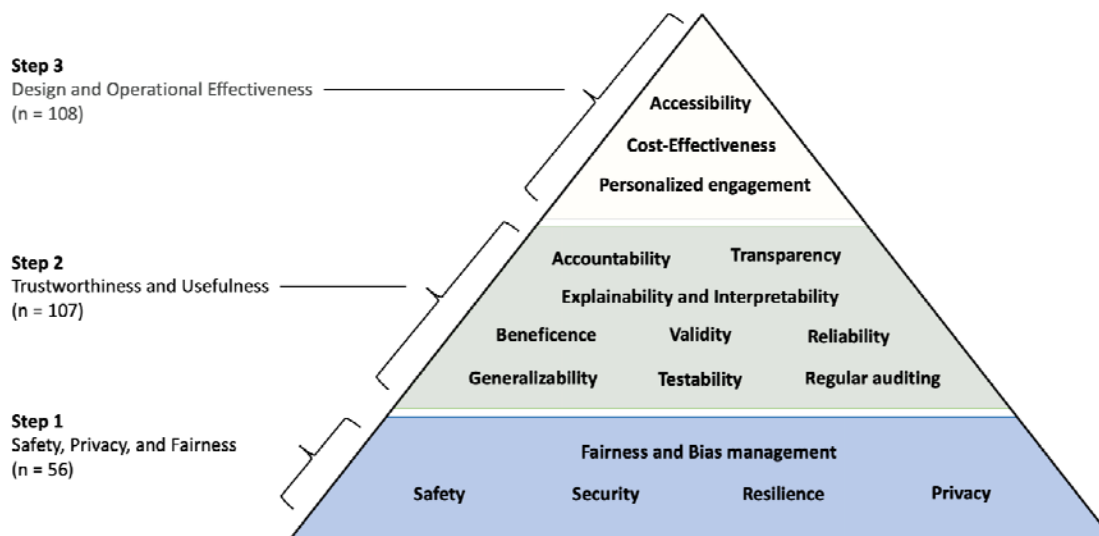
72 As healthcare chatbots face a variety of users, there is no single way to evaluate a chatbot. Factors such as safety  
73 and privacy, user preferences, technology literacy, accessibility, and treatment goals are crucial in determining  
74 the most suitable evaluation method. In addressing these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

85 patients, technology developers, epidemiologists, and policymakers, and 6) further merged and rephrased  
86 questions based on assigned constructs.

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

## 96 Results

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



104 **Figure 1: Pyramid for healthcare chatbot evaluation framework.** Priority-level constructs are displayed on  
105 the left, with second-level constructs within the pyramid.  
106

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

117 The pyramid structure, similar to Maslow's Hierarchy of Needs, serves as a visual reminder that evaluation may  
118 begin at the base, and progression is likely unnecessary if any level fails to meet the required standards. Still, the

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

## 130 **Discussion**

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

136 Our framework's foundation in safety, privacy, and fairness is well aligned with recent research raising concerns  
137 about these aspects of chatbots. A 2024 review of AI apps concluded these apps may cause harm associated with  
138 bias<sup>25</sup> and the 2023 real-world case of an AI chatbot for eating disorders giving dangerous information to users<sup>26</sup>  
139 highlight the importance of Step 1 (see figure 1) in our framework. Not all AI chatbots are patient facing and the  
140 framework is relevant to scaffolding conversations about clinical documentation chatbots, differential diagnosis  
141 chatbots, even scheduling chatbots given the core aspects of the framework are relevant. For example, while  
142 efforts are underway to identify and address bias in conversational agents,<sup>27</sup> checking for and identifying bias in  
143 any chatbot is a productive first step in considering any conversational agent is a foundational step for avoiding  
144 harm.

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

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

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

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

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

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

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

## 188 **Conclusion**

189 This is the first work to develop a structured and adaptable framework for evaluating healthcare AI chatbots,  
190 addressing the urgent need for standardized assessment criteria. By synthesizing insights from existing  
191 frameworks and diverse stakeholders, we developed a structured approach that prioritizes safety, privacy,  
192 trustworthiness, and usefulness. This framework is intended to guide the responsible evaluation and  
193 implementation of chatbots in healthcare, helping to ensure their safe and effective use. Future work will focus  
194 on validating and refining this framework in different contexts.

## 195 **Funding**

196 This study did not receive any funding.

## 197 **Conflict of Interest**

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

## 206 **Declaration of generative AI and AI-assisted technologies in the writing 207 process**

208 During the preparation of this work the author(s) used ChatGPT-4o, web version (accessed 06/26/2024 -  
209 07/02/2024) in order to rephrase some of the framework questions into binary questions (Appendix D). After  
210 using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for  
211 the content of the publication.

## 212 **References**

213 1 Grand View Research. Healthcare Chatbots Market Size, Share & Trends Analysis Report By Component



- 214 (Software, Services), By Application (Appointment Scheduling, Symptom Checking), By Deployment, By  
215 End-user, And Segment Forecasts, 2023 - 2030. San Francisco, CA: Grand View Research, Inc., 2024.
- 216 2 Bach D. How international health care organizations are using bots to help fight COVID-19. *Microsoft* 2020;  
217 published online April. [https://news.microsoft.com/transform/how-international-health-care-organizations-](https://news.microsoft.com/transform/how-international-health-care-organizations-are-using-bots-to-help-fight-covid-19/)  
218 [are-using-bots-to-help-fight-covid-19/](https://news.microsoft.com/transform/how-international-health-care-organizations-are-using-bots-to-help-fight-covid-19/).
- 219 3 Katz U, Cohen E, Shachar E, *et al*. GPT versus Resident Physicians — A Benchmark Based on Official  
220 Board Scores. *NEJM AI* 2024; **1**: AIdbp2300192.
- 221 4 Meaney C, Huang RS, Lu K, Fischer AW, Leung FH, Kulasegaram K, Tzanetos K, Punnett A. Comparing the  
222 performance of ChatGPT and GPT-4 versus a cohort of medical students on an official University of Toronto  
223 undergraduate medical education progress test. medRxiv. 2023 Sep 14:2023-09.
- 224 5 Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-  
225 Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical  
226 education using large language models. *PLoS digital health*. 2023 Feb 9;2(2):e0000198.
- 227 6 Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales N, Tanwani A, Cole-Lewis H, Pfohl S,  
228 Payne P. Large language models encode clinical knowledge. *Nature*. 2023 Aug;620(7972):172-80.
- 229 7. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, Pfeffer MA, Fries J, Shah NH. The shaky  
230 foundations of large language models and foundation models for electronic health records. *npj Digital*  
231 *Medicine*. 2023 Jul 29;6(1):135.
- 232 8 Torous J, Blease C. Generative artificial intelligence in mental health care: potential benefits and current  
233 challenges. *World Psychiatry* 2024; **23**: 1–2.
- 234 9 Organization WH. Ethics and governance of artificial intelligence for health: large multi-modal models.  
235 WHO guidance. World Health Organization, 2024.
- 236 10 Blumenthal D, Patel B. The Regulation of Clinical Artificial Intelligence. *NEJM AI*. 2024 Jul  
237 12:AIpc2400545.
- 238 11. Zhou H, Liu F, Gu B, *et al*. A Survey of Large Language Models in Medicine: Progress, Application, and  
239 Challenge. 2024; published online July 10. DOI:10.48550/arXiv.2311.05112.
- 240 12 Denecke K, Warren J. How to Evaluate Health Applications with Conversational User Interface? *Stud Health*  
241 *Technol Inform* 2020; **270**: 976–80.
- 242 13 Denecke K. Framework for Guiding the Development of High-Quality Conversational Agents in Healthcare.  
243 *Healthcare* 2023; **11**: 1061.
- 244 14 Denecke K, May R. Developing a Technical-Oriented Taxonomy to Define Archetypes of Conversational  
245 Agents in Health Care: Literature Review and Cluster Analysis. *J Med Internet Res* 2023; **25**: e41583.
- 246 15 Liu C, Zowghi D, Peng G, Kong S. Information quality of conversational agents in healthcare. *Inf Dev*  
247 2023; : 02666669231172434.
- 248 16 Martinengo L, Lin X, Jabir AI, *et al*. Conversational Agents in Health Care: Expert Interviews to Inform the  
249 Definition, Classification, and Conceptual Framework. *J Med Internet Res* 2023; **25**: e50767.
- 250 17 Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D’Alfonso S. To chat or bot to chat: Ethical issues  
251 with using chatbots in mental health. *Digit Health* 2023; **9**: 20552076231183542.
- 252 18 Xue J, Zhang B, Zhao Y, *et al*. Evaluation of the Current State of Chatbots for Digital Health: Scoping  
253 Review. *J Med Internet Res* 2023; **25**: e47217.
- 254 19 Abbasian M, Khatibi E, Azimi I, *et al*. Foundation metrics for evaluating effectiveness of healthcare  
255 conversations powered by generative AI. *Npj Digit Med* 2024; **7**: 1–14.
- 256 20 Shlobin NA, Ward M, Shah HA, *et al*. Ethical Incorporation of Artificial Intelligence into Neurosurgery: A  
257 Generative Pretrained Transformer Chatbot-Based, Human-Modified Approach. *World Neurosurg* 2024;  
258 **187**: e769–91.
- 259 21 Ding H, Simmich J, Vaezipour A, Andrews N, Russell T. Evaluation framework for conversational agents  
260 with artificial intelligence in health interventions: a systematic scoping review. *J Am Med Inform Assoc*  
261 2024; **31**: 746–61.
- 262 22 Nadarzynski T, Knights N, Husbands D, *et al*. Achieving health equity through conversational AI: A  
263 roadmap for design and implementation of inclusive chatbots in healthcare. *PLOS Digit Health* 2024; **3**:  
264 e0000492.
- 265 23 Henson P, David G, Albright K, Torous J. Deriving a practical framework for the evaluation of health apps.  
266 *Lancet Digit Health* 2019; **1**: e52–4.
- 267 24 Coalition for Health AI. Blueprint for Trustworthy AI: Implementation Guidance and Assurance for  
268 Healthcare. 2023; published online April. <https://www.coalitionforhealthai.org/>.
- 269 25 Wongvibulsin S, Yan MJ, Pahalyants V, Murphy W, Daneshjou R, Rotemberg V. Current State of  
270 Dermatology Mobile Applications With Artificial Intelligence Features. *JAMA Dermatol* 2024; **160**: 646–50.
- 271 26 Sharp G, Torous J, West ML. Ethical Challenges in AI Approaches to Eating Disorders. *J Med Internet Res*  
272 2023; **25**: e50696.
- 273 27 Flores L, Kim S, Young SD. Addressing bias in artificial intelligence for public health surveillance. *J Med*

- 274 *Ethics* 2024; **50**: 190–4.  
275 28 King M. Harmful biases in artificial intelligence. *Lancet Psychiatry* 2022; **9**: e48.  
276 29 De Freitas J, Cohen IG. The health risks of generative AI-based wellness apps. *Nat Med* 2024; **30**: 1269–75.  
277 30 Hua Y, Liu F, Yang K, *et al.* Large Language Models in Mental Health Care: a Scoping Review. 2024;  
278 published online Jan 1. DOI:10.48550/arXiv.2401.02984.  
279 31 Chen S, Kann BH, Foote MB, *et al.* Use of Artificial Intelligence Chatbots for Cancer Treatment Information.  
280 *JAMA Oncol* 2023; **9**: 1459–62.

## 281 **Appendix A: Search Strategy**

### 282 **A.1. Search Strategy**

283 To identify and evaluate existing frameworks for healthcare conversational agents, we followed the PRISMA  
284 guidelines to conduct a systematic review. The literature search was performed across multiple databases to  
285 ensure comprehensive coverage of relevant studies. The databases and corresponding search terms were as  
286 follows:

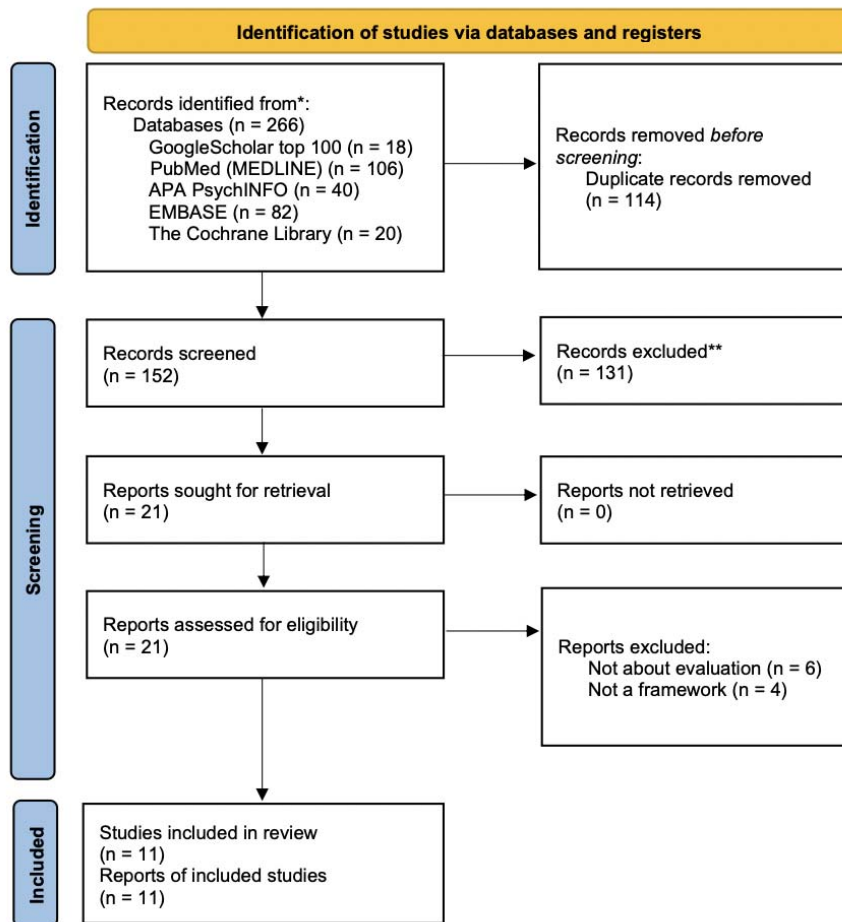
Database	Query
PubMed (MEDLINE)	("health"[Title/Abstract] OR "medical"[Title/Abstract] OR "medicine"[Title/Abstract] OR "clinical"[Title/Abstract]) AND ("conversational agent"[Title/Abstract] OR "conversational AI"[Title/Abstract] OR "chatbot"[Title/Abstract] OR "virtual agent"[Title/Abstract] OR "virtual assistant"[Title/Abstract] OR "digital assistant"[Title/Abstract]) AND ("framework"[Title/Abstract] OR "evaluation method"[Title/Abstract] OR "assessment method"[Title/Abstract])
EMBASE	('health':ti,ab OR 'medical':ti,ab OR 'medicine':ti,ab OR 'clinical':ti,ab) AND ('conversational agent':ti,ab OR 'conversational ai':ti,ab OR 'chatbot':ti,ab OR 'virtual agent':ti,ab OR 'virtual assistant':ti,ab OR 'digital assistant':ti,ab) AND ('framework':ti,ab OR 'evaluation method':ti,ab OR 'assessment method':ti,ab)
APA PsychINFO	("health" OR "medical" OR "medicine" OR "clinical") AND ("conversational agent" OR "conversational AI" OR "chatbot" OR "virtual agent" OR "virtual assistant" OR "digital assistant") AND ("framework" OR "evaluation method" OR "assessment method")
The Cochrane Library	("health" OR "medical" OR "medicine" OR "clinical") AND ("conversational agent" OR "conversational AI" OR "chatbot" OR "virtual agent" OR "virtual assistant" OR "digital assistant") AND ("framework" OR "evaluation method" OR "assessment method")
Google Scholar*	("health" OR "medical" OR "clinical") AND ("conversational agent" OR "conversational AI" OR "chatbot" OR "virtual agent" OR "virtual assistant" OR "digital assistant") AND ("framework" OR "evaluation method" OR "assessment method")

287  
288 \*Note: Since Google Scholar does not support advanced search queries, we performed all combinations of  
289 searches separately to ensure comprehensive coverage.

### 290 **A.2. Inclusion and Exclusion Criteria**

291 The search was restricted to full-length papers published between January 1, 2018, and June 25, 2024. We  
292 included studies that developed frameworks for evaluating healthcare conversational agents. We excluded  
293 studies introducing new evaluation methods without the intention of providing a structural evaluation  
294 framework, such as clinical trials and model development studies.

### 295 **A.3. Screening and Selection Process**



296

297 **Figure A.3.** PRISMA Flow Diagram of Study Selection for Evaluation Frameworks of Healthcare  
298 Conversational Agents.

299

300 The initial search results were screened based on titles and abstracts. Two authors (YH and WX) independently  
301 reviewed the titles and abstracts for full-text retrieval, with any discrepancies resolved by discussion with a third  
302 reviewer (JT). Full-text articles were then retrieved for further assessment against the inclusion criteria. YH  
303 reviewed the full texts and verified them with JT. From the initial 266 records, 152 were screened, and 21  
304 reports were sought for retrieval. After detailed assessment, 11 studies were included in the review, providing a  
305 comprehensive evaluation of frameworks for healthcare conversational agents.

306 **Appendix B: Reviewed Frameworks**

Title	Year	Term Used for CA	Intention
How to Evaluate Health Applications with Conversational User Interface?	2020	Conversational User Interface (CUI), Chatbot	Support evaluation of health systems using CUIs, define quality dimensions, guide developers and researchers.
Conversational Agents in Health Care: Expert Interviews to Inform the Definition, Classification, and Conceptual Framework	2023	Conversational Agent	Define and classify health care CAs, validate the DISCOVER conceptual framework, update CHAT framework focusing on ethics, user involvement, and data privacy.
Developing a Technical-Oriented Taxonomy to Define Archetypes	2023	Conversational Agent	Develop taxonomy of technical characteristics, identify archetypes,



of Conversational Agents in Health Care: Literature Review and Cluster Analysis			harmonize evaluation metrics.
Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review	2023	Conversational Agent	Propose a four-stage evaluation framework (feasibility/usability, efficacy, effectiveness, implementation) based on WHO recommendations.
Evaluation of the Current State of Chatbots for Digital Health: Scoping Review	2023	Chatbot	Assess current state of health-related chatbots, identify research gaps, guide future research, and enhance chatbot design.
Framework for Guiding the Development of High-Quality Conversational Agents in Healthcare	2023	Conversational Agent	Provide a framework for the development and evaluation of health CAs, ensure patient safety, and efficacy of CA-delivered interventions.
Information quality of conversational agents in healthcare	2023	Conversational Agent	Investigate definitions, influencing factors, and impacts of information quality (IQ) in health CAs.
To chat or bot to chat: Ethical issues with using chatbots in mental health	2023	Chatbot	Examine ethical issues in using chatbots in mental health, provide recommendations for ethical design and deployment.
Ethical Incorporation of Artificial Intelligence into Neurosurgery: A Generative Pretrained Transformer Chatbot-Based, Human-Modified Approach	2024	Chatbot, Generative Pretrained Transformer (GPT)	Delineate ethical considerations for AI in neurosurgery, present an ethical framework for AI integration.
Achieving health equity through conversational AI: A roadmap for design and implementation of inclusive chatbots in healthcare	2024	Conversational AI, Chatbot	Develop a roadmap for inclusive conversational AI in healthcare, promote health equity.
Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI	2024	Conversational AI, Large Language Models (LLMs)	Establish a framework for evaluating effectiveness of healthcare conversations using generative AI, address limitations of existing metrics.

307

### 308 **Appendix C: Details on the review process**

309 We began by summarizing each framework's intended use to assess specific concepts within a particular domain.  
 310 The sections detailing the evaluation framework's questions were then extracted and listed. If the study did not  
 311 explicitly present evaluation criteria in the form of questions, these criteria were rephrased as questions for  
 312 clarity. The following steps were taken:

- 313 • Describe Use Intention: The purpose and intended application of the framework were articulated,  
 314 highlighting its relevance and scope.
- 315 • Concepts Evaluated: The key concepts and dimensions the framework evaluates were identified and  
 316 outlined.
- 317 • Listing Evaluation Questions: A thorough list of the questions evaluated by the framework was  
 318 provided. In cases where the study did not present evaluation criteria as questions, these criteria were  
 319 rephrased into question format for consistency and clarity.

320

321 Initially, we broke down questions that contained multiple sub questions. Questions too broad to be constructive  
 322 were then removed. For instance, we did not include questions such as: "Can strategies or solutions be  
 323 developed to address problems of CAs?" and "Does the AI system comply with national and international  
 324 regulations and standards?"

## 325 **Appendix D: Extracted Questions and Final Questions**

326 Find in the downloadable supplementary file.

## 327 **Appendix E: Tree-structured Framework**

### 328 **1. Safety, privacy, and fairness:**

- 329 **a. Safety:** prevent worse outcomes for the patient, provider, or health system from occurring as a  
330 result of the use of an ML algorithm.
- 331 **i. Outcome proxies appropriateness:** use alternative measures or indicators that accurately  
332 reflect the desired health outcomes in the absence of direct measurements.
  - 333 **ii. Data provenance:** track and document the origin and history of data, including where it  
334 came from and how it has been handled.
    - 335 **1. Data Providers:** assign roles and responsibilities to entities like hospital EHRs  
336 and patient-generated health data for maintaining safe AI.
    - 337 **2. Data Sources:** include various origins of data such as social media and clinical  
338 settings.
  - 339 **iii. Harm control:** reduce and manage potential risks and negative impacts associated with  
340 using a chatbot.
  - 341 **iv. Reducing automation bias** (i.e., the tendency to accept automated suggestions without  
342 critical evaluation or questioning)
  - 343 **v. Critical help:** provide necessary assistance and address negative and help-seeking  
344 information.
  - 345 **vi. Ethics:** principles and standards that govern the conduct of individuals and organizations,  
346 ensuring fairness, privacy, and respect in using ML algorithms in healthcare.
- 347 **b. Security:** maintain confidentiality, integrity, and availability through protection mechanisms that  
348 prevent unauthorized access and use
- 349 **i. Protection method:** implement techniques and tools to safeguard data from unauthorized  
350 access and threats.
  - 351 **ii. Security standard:** follow established guidelines and practices designed to protect data  
352 and systems from security breaches.
  - 353 **iii. Third-party reliability:** ensure the trustworthiness of external partners or services in  
354 maintaining data security and integrity.
- 355 **c. Resilience:** withstand unexpected adverse events or changes in their environment or use
- 356 **d. Privacy:** protect privacy according to standards like HIPAA and GDPR, ensuring user autonomy  
357 and dignity.
- 358 **i. Data exchange:** maintain privacy standards for accessing and sharing data with third-  
359 party tools, cloud platforms, and other external systems.
  - 360 **ii. Data collection and storage:** maintain privacy standards for gathering and securely  
361 storing data for future use.
  - 362 **iii. Data usage:** maintain privacy standards for using collected data for analysis, decision-  
363 making, and improving chatbot algorithms.
  - 364 **iv. Privacy Policy:** outline how an organization collects, uses, protects, and shares personal  
365 data.
  - 366 **v. Data protection:** implement methods to ensure privacy and prevent unauthorized access  
367 and breaches.
- 368 **e. Fairness and Bias Management:** ensure the chatbots operate with minimized and acknowledged  
369 biases to ensure fair outcomes.
- 370 **i. Systemic Bias:** address biases originating from societal norms and institutional practices.
  - 371 **ii. Computational and Statistical Bias:** manage biases arising from the way data is  
372 processed and algorithms are designed.
  - 373 **iii. Human-cognitive biases:** recognize biases stemming from individual or group  
374 perceptions and attitudes.
  - 375 **iv. Population bias:** address the issue where certain populations are underrepresented in data,  
376 leading to less accurate model performance for those groups.
- 377 **2. Trustworthiness and Usefulness**
- 378 **a. Accountability:** ensure those involved in the chatbot's lifecycle uphold standards of auditability  
379 and harm minimization.
  - 380 **b. Transparency:** communicate clearly regarding the chatbot's characteristics and performance  
381 throughout its lifecycle.
    - 382 **i. Usage Specification:** define how the chatbot should be used.

- 383           ii. **Model Characteristics:** describe the specific features and behaviors of the chatbot.
- 384           iii. **Model Availability:** ensure the chatbot is accessible as needed.
- 385           iv. **Model Limitations:** identify and communicate the boundaries and constraints of the
- 386 chatbot.
- 387           v. **Data Usage:** explain how data is utilized within the chatbot.
- 388       c. **Explainability and interpretability:**
- 389           i. **Model Explainability:** detail the internal mechanisms and decision-making processes of
- 390 the chatbot.
- 391           ii. **Model Interpretability:** make the outputs of chatbots clear and meaningful to end-users.
- 392       d. **Beneficence:** ensure chatbot positively impacts its intended outcomes, emphasizing measurable
- 393 benefits over potential risks.
- 394           i. **Health Outcomes:** focus on improving health results.
- 395           ii. **Clinical Evidence:** use rigorous methods like A/B tests or RCTs to validate effectiveness.
- 396           iii. **User Behaviors:** influence and improve user actions.
- 397           iv. **Intervention:** apply targeted measures to achieve desired outcomes.
- 398           v. **Healthcare System:** integrate effectively within the broader healthcare environment.
- 399       e. **Validity:** ensure the chatbot performs as expected in real-world conditions.
- 400           i. **Data Relevance and Credibility:** use high-quality, pertinent training data.
- 401           ii. **Language Understanding:** ensure the chatbot's linguistic capabilities are robust.
- 402           iii. **Information Retrieval Accuracy:** accurately retrieve relevant information.
- 403           iv. **Outcome Accuracy:** deliver precise and correct results.
- 404           v. **Task Completion:** effectively complete required functions.
- 405       f. **Reliability:** ensure that the chatbot consistently performs as intended under various conditions
- 406 and maintains dependable operation over time.
- 407           i. **Failure Prevention:** prevent system failures to maintain functionality.
- 408           ii. **Robustness:** handle unexpected inputs and diverse data without errors.
- 409           iii. **Workflow Integration:** fit seamlessly into existing processes.
- 410           iv. **Reproducibility:** ensure consistent outcomes across different settings.
- 411           v. **Monitoring:** continually check chatbots to ensure proper operation.
- 412           vi. **Up-to-dateness:** keep the system current with the latest information.
- 413       g. **Generalizability:** apply learned patterns to new, unseen data.
- 414           i. **Contextual Adaptability:** function effectively in different environments or clinical
- 415 contexts.
- 416               1. **Age Group Adaptability:** cater to different age groups.
- 417               2. **Scenario Adaptability:** adapt to various situations.
- 418           ii. **Novel Data Performance:** perform well with new, unseen data.
- 419       h. **Testability:** verify and meet standards for robustness, safety, bias mitigation, fairness, and equity.
- 420           i. **Verifiability:** ensure different attributes can be tested.
- 421               1. **Quantifiability:** measure attributes precisely.
- 422           ii. **Regular Auditing:** measure attributes regularly.
- 423       3. **Design and Operational Effectiveness**
- 424           a. **Accessibility:** ensure the chatbot is usable by the intended users, regardless of their abilities,
- 425 devices, or technical skills, promoting inclusivity and ease of use.
- 426           i. **Versatile access:** provide multiple interaction methods to accommodate user preferences
- 427 and needs.
- 428               1. **Multi-language:** enable interaction in multiple languages to cater to a diverse
- 429 user base.
- 430               2. **Different Input and Output Mode:** accommodate various input and output
- 431 methods, such as text, voice, and visual.
- 432               3. **Multi-platform:** ensure functionality across different platforms, such as web,
- 433 mobile, and desktop applications.
- 434               4. **Multi-device:** provide compatibility with various devices, including
- 435 smartphones, tablets, laptops, and desktop computers.
- 436           ii. **User literacy:** ensure the system is usable by individuals with varying levels of technical
- 437 knowledge and literacy.
- 438           iii. **User experience:** create a pleasant and effective interaction for users.
- 439               1. **Likability:** design the system to be appealing and enjoyable to use.
- 440               2. **Understood by the CA (Conversational Agent):** ensure clear communication
- 441 between the user and the chatbot.
- 442               3. **User Engagement:** maintain user interest and active participation.

- 443                                   **4. Respectfulness:** interact with users in a polite and respectful manner.
- 444                                   **5. Response Appropriateness:** provide suitable and contextually relevant
- 445                                   responses.
- 446                                   **6. Credibility:** ensure the chatbot's reliability and trustworthiness.
- 447                   **iv. User Interface Design:** create an intuitive and easy-to-use interface for users.
- 448                   **v. Simplicity/Ease of Use:** make the system straightforward and user-friendly, minimizing
- 449                   complexity and effort required from users.
- 450           **b. Personalized engagement:** tailor responses based on patient data and preferences.
- 451                   **i. Personalization:** customized response based on patient data and preference
- 452                   **ii. Anthropomorphism/relationship:** build a human-like relationship with users.
- 453                                   **1. Relationship Building:** develop a rapport with users.
- 454                                   **2. Empathy:** show understanding and compassion.
- 455                                   **3. Humor:** use appropriate humor to engage users.
- 456                                   **4. Identity:** establish a clear and consistent chatbot persona.
- 457                   **iii. User Adherence:** track and analyze how well users follow recommendations, and adjust
- 458                   the chatbot's strategies based on this data to improve compliance and outcomes
- 459                   **iv. Feedback Incorporation:** use user feedback to improve the system.
- 460                   **v. Progress awareness:** monitor and respond to the conversation's context and progress.
- 461                                   **1. Memory:** support multi-turn or multi-session conversations.
- 462                                   **2. Strategy Adjustment:** adapt the conversation strategy as needed.
- 463           **c. Cost-effectiveness:** assess whether the chatbot delivers beneficial outcomes at a reasonable cost,
- 464           providing a better or more economical solution compared to existing methods.
- 465                   **i. Comparative Effectiveness:** demonstrate that the chatbot is a better solution than
- 466                   previous methods.
- 467                   **ii. Economical Viability:** ensure the system is cost-effective.
- 468                   **iii. Environmental Viability:** minimize environmental impact.
- 469                   **iv. Task Efficiency:** perform tasks quickly and effectively.
- 470                                   **1. Appropriate Response Time:** provide timely responses.
- 471                                   **2. Response Conciseness:** give clear and succinct information.
- 472                                   **3. Response Relevance:** ensure responses are pertinent to the query.
- 473                                   **4. Response Practicality:** offer practical and actionable information.
- 474                   **v. Workflow Considerations:** integrate smoothly into existing systems.
- 475

476 Questions under constructs such as accessibility assurance and accountability assurance (referenced in Appendix  
477 C - Final Questions and Appendix F - Framework Questions Statistics ) only assess whether their parent  
478 constructs (accessibility and accountability, respectively in this case) are ensured in the evaluation. These  
479 placeholder-like subconstructs are not included in this framework for simplicity. Further work is needed to  
480 develop questions and future classifications for these constructs, as they are currently overlooked by the  
481 literature.

482 **Appendix F: Framework Question Statistics**

