International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct 2026, PP 01-7

AI-Driven Chatbots in Public Health: A Review of Applications, Evidence, and Challenges

Satvik V. Sumbare, Prof. Bangar A. P., Prof. Bhosale S.B., Dr. A. A. Khatri
Dept. of Computer Engineering,
Jaihind collage of Engineering, Pune, India

Abstract. Chatbots driven by artificial intelligence (AI) are increasingly used to deliver health information, screening, counselling, and behavioural support at population scale. This paper reviews five recent secondary studies, namely an umbrella review of 44 systematic reviews, a systematic review of large language model (LLM) chatbots, and three scoping reviews of COVID-19 and public health chatbots, to consolidate what is known about how these tools are designed, where they are applied, and how well they work. Most deployed chatbots are text-based, rely on simple decision-tree or hybrid architectures, and are concentrated in pandemic response, mental health, and health-information provision. Evidence of effectiveness is encouraging for smoking cessation and for short-term reductions in depression and anxiety, but mixed for lifestyle change and clinical decision support, and benefits often fade at follow-up. Persistent weaknesses include a shortage of randomised trials, limited use of theory and user-centred design, inconsistent reporting, readability and accuracy problems in LLM output, slow safety escalation, privacy risks, and digital-divide inequities. The paper closes with priorities for standardised evaluation, domain-specific model development, safeguarding, and inclusive design.

Keywords: chatbots, conversational agents, artificial intelligence, public health, large language models, digital health, mental health, COVID-19

I. Introduction

Chatbots are software programs that converse with people in natural language through text or speech. They have moved from early experiments such as ELIZA to routine components of digital health services [5]. Conversational agents are commonly described as interactive programs that mimic human interaction using text, voice, or other non-verbal cues [1], [6]. Their appeal in public health is practical: they are available at any hour, can talk with very large numbers of people at once, can be reached through apps that people already use, and can relieve staff of repetitive questions.

The COVID-19 pandemic accelerated adoption. The World Health Organization (WHO) Health Alert chatbot on WhatsApp reportedly reached more than 12 million people within its first month of operation [3], and AI-supported symptom checking and triage became commonplace around the world [5]. In parallel, generative AI, and in particular large language models (LLMs) made widely available from late 2022, changed what a chatbot can do by letting it compose free-form, context-sensitive replies instead of selecting scripted ones [2].

Despite this growth, the evidence is scattered across clinical domains, technologies, and study designs, and individual reviews often reach different conclusions about effectiveness [1]. This paper therefore synthesises five recent secondary studies to address four questions: (i) how are AI-driven chatbots designed and deployed in public health; (ii) which applications are most

common; (iii) what does the evidence say about effectiveness, and how is it evaluated; and (iv) which challenges and research priorities emerge. The contribution is a single consolidated view that links design choices, applications, outcomes, and risks.

II. Review Approach

This is a narrative review of five peer-reviewed secondary studies chosen for their complementary scope (Table I). Together they cover pandemic-era chatbots, public health chatbots in general, LLM-based chatbots, and a broad umbrella review of 44 systematic reviews published between 2000 and 2025. For each study, information was extracted on chatbot type, technology, platform, application domain, reported outcomes, evaluation methods, and stated limitations, and the material was then organised thematically.

Two cautions apply. First, the reviews draw on overlapping primary studies, so counts from different reviews are not additive. Second, figures are reported as stated by the original authors; the underlying primary studies were not re-appraised here.

Table 1: Secondary Studies Included in This Review

Ref.	Design and scope	Principal contribution
[3]	Scoping review of COVID-19 chatbots; 9 of 543 articles	Five application types; most papers describe development, only three study users
[4]	Scoping review; 61 chatbots in more than 30 countries	Six use-case categories and 15 use cases; mostly simple decision-tree designs
[5]	Scoping review of public health chatbots; 32 of 1,506 studies	Mental health and COVID-19 dominate; efficacy evidence limited and mixed
[1]	Umbrella review; 44 systematic reviews, 2000–2025	Broad map of functions; outcomes inconsistent across domains
[2]	Systematic review of LLM-based chatbots; 20 of 348 articles, 2022–2024	Four research areas: text quality, clinical efficacy, engagement, safety

III. Chatbot Technology and Design

1. From Rule-Based Systems to LLMs

Rule-based chatbots follow predefined expert rules, whereas AI-based chatbots use machine learning, deep learning, or LLMs to interpret input and generate replies [2]. In the COVID-19 literature, answers were typically built from question-and-answer pairs drawn from official FAQs using Artificial Intelligence Markup Language [3]. An analysis of 61 pandemic chatbots found that most were relatively simple, used decision-tree structures with predetermined response options, and focused on a narrow set of tasks, presumably because they had to be deployed quickly [4].

Among the reviews in the umbrella study that described their methods, deep learning was the most commonly reported technique (14 reviews), usually combined with natural language processing [1]. The LLM-focused review shows a spectrum rather than a clean divide. QuitBot pairs a structured 42-day cessation programme and a scripted question library with a fine-tuned GPT-3.5 model for out-of-library questions. Wysa uses language models to select or phrase replies inside a clinician-vetted

International Journal of Statistics and Mathematical Research Trends
E-ISSN: 3108-0871, Volume 2 Issue 5
Sep-Oct 2026, PP 01-7

library with safety guardrails. The WHO Florence chatbot and the Mind Tutor app use AI only to interpret user input and do not generate free text [2]. Such hybrid designs trade conversational flexibility for predictability and safety.

2. Interaction Modality and Platform

Text is the dominant modality. Of the pandemic chatbots whose interface could be assessed, 42 were text-based, 4 voice-based, and 4 offered both [4]. The umbrella review similarly found text-based chatbots most common (18 reviews), followed by voice-based (15), embodied (10), and combined text-and-voice agents (6) [1]. Only 4 of 32 chatbots in the public health scoping review replied in speech [5].

Deployment platforms are varied. Among the pandemic chatbots, 34 were web-based, 22 on social media, 8 standalone or health apps, 6 on SMS, and 4 telephone-based, and 30 were embedded in an organisation's own platform to improve visibility [4]. The umbrella review reported stand-alone or embedded applications (13 reviews), web (8), social media (7), SMS (7), and virtual reality (7) [1]. Messaging-app and built-in bot tools speed development and use an existing user base, while SMS and phone channels help reach people without smartphones or reliable internet [4].

3. Personas and Design Process

Most pandemic chatbots (70%) lacked anthropomorphic features, and where gender was given it was mostly female [4]. In the public health scoping review, 17 chatbots used a static avatar, 8 of them human-like, and the five embodied agents were all presented as female [5]. The influence of persona on outcomes remains under-researched, although QuitBot's empathetic, socially engaging persona was associated with stronger user connection [2]. Systematic design methods were uncommon: only 3 of 32 studies described a user-centred design process, with the community-designed contraception chatbot Layla as a notable exception [5]. Only three pandemic chatbots initiated follow-up contact [4].

IV. Applications in Public Health

1. Pandemic Response

The most systematic mapping of use cases comes from a review of 61 chatbots deployed in more than 30 countries [4]. Six categories emerged (Fig. 1): information dissemination (42 chatbots), risk assessment (34), surveillance (12), post-COVID eligibility screening (2), distributed coordination (1), and a vaccine scheduler (1). A smaller scoping review of nine studies independently identified five applications: health information (77.7% of studies), self-triage and risk assessment (66.6%), exposure monitoring and notification (55.5%), symptom tracking (44.4%), and countering misinformation (33.3%) [3].

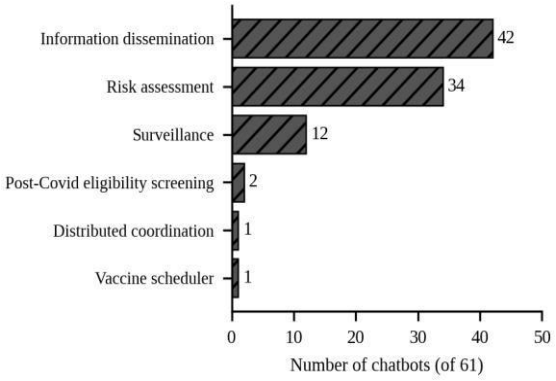


Fig.1. Use-case categories of 61 COVID-19 public health chatbots (data from [4]; a chatbot may serve more than one category).

Risk-assessment bots triage users by symptoms and exposure and recommend a course of action, which reduces footfall in healthcare facilities and helps free capacity [4]. Surveillance bots collect prevalence or symptom data and, in some cases, monitor high-risk groups proactively; Japan's COOPERA system built on the LINE messenger is one example [3], [4]. One distributed-coordination platform, mHero, connects frontline health workers with ministries of health in settings with weak infrastructure, which is the only case in the sample directed at staff rather than the public [4].

2. Mental Health Support

Mental health is the most studied clinical area. In the umbrella review, 23 reviews covered mental health support, concentrated on depression and anxiety, with few reviews addressing conditions such as ADHD or bipolar disorder [1]. In the public health scoping review, mental health accounted for 10 of 32 studies, the largest single domain [5], and in the LLM-focused review acting as a personal coach for mental health counselling was one of the two leading functions (40% of studies) [2]. Typical products deliver cognitive behavioural therapy techniques, mood tracking, and self-help guidance, and they are best positioned as supplements to, not replacements for, human care [5].

3. Behaviour Change and Addiction

Twelve reviews examined chatbots for healthy lifestyle (physical activity, diet, weight, sleep) and ten for addiction, mainly smoking cessation and substance use [1]. Chatbots suit these tasks because behaviour change depends on repeated, timely prompts, personalised feedback, and low-barrier access, all of which can be automated at scale.

4. Triage, Decision Support, and Counselling

Clinical decision support was the most frequently reported function in the umbrella review (29 reviews), including assessment of clinical vignettes or diagnosis (15) and consultation or counselling (19) [1]. During the pandemic, chatbots screened symptoms and directed users toward testing or care [4], [5]. With the arrival of LLMs, many recent studies have asked clinicians to rate the quality of chatbot answers to patient questions across specialties such as surgery, radiology, and addiction medicine [2].

5. Education, Self-Management, and Administration

Twenty reviews covered educational uses, and 18 reported symptom tracking and self-management of disease [1]. Less common but promising tasks include patient intake and workflow automation, medication adherence support, appointment booking, and accessible health-data entry for older adults and people with sensory impairments [1], [2]. Table II summarises the applications and the status of their evidence.

Table 2: Application Domains and Evidence Status

Domain	Evidence status (sources)
Pandemic response	Widely deployed; mostly descriptive evidence; symptom-checker accuracy varied [4], [5]
Mental health	Most studied; short-term benefits common, long-term often not significant [1]
Smoking and substance use	Most consistent positive results [1], [2]
Lifestyle and weight	Mixed or inconclusive [1], [5]
Decision support and triage	Mixed; accuracy depends on condition and model [1], [5]
Education and administration	Promising, thinly evaluated [1], [2]

V. Evidence of Effectiveness

1. Encouraging Results

The strongest signal is for smoking cessation. In the umbrella review, the five reviews that reported the direction of addiction-related outcomes were all positive [1], and a meta-analysis found that users of conversational AI interventions were significantly more likely to quit [9]. In the LLM-focused review, the QuitBot randomised controlled trial (RCT) showed higher quit rates [2].

For mental health, 13 reviews judged chatbots effective for improving or managing symptoms, whereas seven found insignificant or mixed effects [1]. Meta-analyses reported moderate to large effects on depression (Hedges' g= 0.64) and distress (g = 0.7) [11], small to moderate short-term effects on depression (g = 0.29) and anxiety (g = 0.29) [10], and modest improvements in depression (g = −0.26) and anxiety (g = −0.19), strongest around eight weeks [1]. Across the public health scoping review, results were generally positive or mixed, and about two-thirds of the studies showed that the chatbots were usable [5]. An earlier systematic review likewise found positive or mixed evidence in 77% of studies [8].

2. Mixed and Weak Results

Benefits often fade. Long-term effects on generalised and specific anxiety were not significant in one meta-analysis [10], and no substantial effects remained at the three-month follow-up in another [1]. Overall well-being did not improve significantly in a further review (g = 0.32, confidence interval spanning zero) [1]. The Mind Tutor RCT found no efficacy [2].

Lifestyle outcomes are inconsistent: some reviews reported weight loss and diet or activity gains, while others found differences from control groups to be insignificant or called the evidence inconclusive [1]. Behaviour-change effects were limited to users already contemplating change, a gambling trial found no significant effect, and adherence results were split [5]. For decision support, a comparison found ChatGPT-3 correct in 83.3% of differential diagnoses versus 98.3% for physicians, and 53.3% versus 93.3% for principal diagnoses; reported accuracy of conversational agents ranged from 14% to 100% depending on the clinical domain [1]. A comparison of ten web symptom checkers found only two with a good balance between sensitivity and specificity [5].

3. How the Evidence Is Generated

Study designs are dominated by quantitative methods (26 of 32 studies), with 12 RCTs and many user-analytics studies [5]. Among the 20 LLM-focused studies, only two were RCTs and only one used a theoretical framework to assess engagement [2]. Studies of text quality typically had clinicians grade responses for accuracy and readability. None of the 32 public health studies used direct observation or in-depth interviews [5], so little is known about how people actually experience these tools. User engagement is a recurring theme. Engagement was associated with a strong persona, the ability to ask free-form questions, and real-time support, and it was reduced by misunderstood input, repetitive or inconsistent replies, interface problems, and perceived privacy risk [2].

Challenges and Risks

- Evidence quality and reproducibility: Many studies use small, short, or homogeneous samples, and 28 of 32 public health studies gave very little technical description of the chatbot, which prevents replication [5]. LLM behaviour can change between studies because of unannounced model updates and probabilistic sampling, which undermines repeatability [2]. Heterogeneity of methods and outcomes also limits cross-study conclusions [1].
- Accuracy, readability, and bias: In the LLM review, most studies found chatbot text accurate and evidence-based, and over 90% of ChatGPT responses in two studies closely matched clinical guidelines, but readability was often poor and the best-performing model differed between studies [2]. One review of oncology chatbots reported bias or hallucination

in 41.7% of included articles [1]. Reviews of LLM applications in health care call for testing with real patient-care data [12].

- Safety: Safety was examined in only one of 20 LLM studies, which found that a GPT-3.5 chatbot was slow to escalate a simulated suicide-risk scenario and that the eventual shutdown came from software guardrails rather than the model itself [2]. Few public health studies described safeguarding at all; where they did, the chatbot simply told users to contact official services [5]. Empathetic language generation is still immature, so chatbots may be unsuitable for people in acute distress [3].
- Privacy and ethics: Perceived privacy risk was negatively associated with engagement [2]. Government use of chatbot and contact-tracing data raised human-rights concerns in at least one national response [5], and WHO has issued guidance on the ethics and governance of AI for health [13]. Chatbots that need identifying information for access cannot be anonymous, whereas follow-up of symptomatic users requires exactly such data, creating a tension between privacy and functionality [4].
- Equity and acceptance: People without internet access or digital skills may be excluded, which can bias surveillance data and widen disparities [3], [4]. Some health chatbots became inactive because of low community interest, and in one experiment users rated a screening chatbot lower than a human agent on ability, integrity, and benevolence [3]. Evaluation samples are often small and geographically narrow [2].
- Keeping knowledge current: Guidance changed daily during the pandemic, and fact-checking while deployed was hard; automating extraction of question-and-answer pairs from trusted sources is one proposed remedy [3], [4].

Future Directions

Several priorities follow from the reviewed work. First, the field needs standardised reporting of chatbot architecture, content, and model versions, together with common evaluation criteria, so that studies can be compared and reproduced [2], [5]. Second, larger RCTs in diverse cultural and demographic settings with longer follow-up are needed to test whether short-term benefits persist [2], [10]. Third, domain-specific models trained on real clinical data, and approaches that keep scripted, clinician-vetted content for high-risk topics, may balance flexibility with safety [2], [12].

Fourth, safety and governance deserve research of their own, including timely escalation to human support, monitoring after deployment, and privacy-preserving designs [2], [13]. Fifth, participatory and user-centred design, combined with qualitative methods such as observation and interviews, should become routine, especially for populations at risk of exclusion [5]. Finally, understudied areas deserve attention, including medication adherence, rehabilitation, conditions such as ADHD and bipolar disorder, support for frontline health workers, and proactive follow-up [1], [4].

VI. Conclusion

AI-driven chatbots have already become a working part of public health practice, valued for scale, accessibility, and the ability to automate repetitive communication. The reviewed literature supports cautious optimism: the best evidence concerns smoking cessation and short-term relief of depressive and anxiety symptoms, while evidence for lifestyle change and diagnostic support is inconsistent. The pace of deployment has outrun the pace of evaluation. Closing that gap through rigorous trials, transparent reporting, safety-first design, and inclusive engagement will determine whether chatbots complement human-led care and extend it fairly to those who need it most.

Acknowledgment

The authors acknowledge the researchers whose published reviews formed the basis of this work.

References

1. A. L. Huynh, T. J. Roy, K. N. Jackson, A. G. Lee, W. Liaw, and M. M. Hossain, "Applications of artificial intelligence-based conversational agents in healthcare: A systematic umbrella review," Int. J. Med. Inform., vol. 207, Art. no. 106204, 2026, doi: 10.1016/j.ijmedinf.2025.106204.
2. S. Feng, X. Li, and A. N. Wake, "Engaging artificial intelligence (AI)-based chatbots in digital health: A systematic review," PLOS Digit. Health, vol. 5, no. 2, e0001201, 2026, doi: 10.1371/journal.pdig.0001201.
3. M. Almalki and F. Azeez, "Health chatbots for fighting COVID-19: A scoping review," Acta Inform. Med., vol. 28, no. 4, pp. 241–247, 2020, doi: 10.5455/aim.2020.28.241-247.
4. P. Amiri and E. Karahanna, "Chatbot use cases in the Covid-19 public health response," J. Am. Med. Inform. Assoc., vol. 29, no. 5, pp. 1000–1010, 2022, doi: 10.1093/jamia/ocac014.
5. L. Wilson and M. Marasoiu, "The development and use of chatbots in public health: Scoping review," JMIR Hum. Factors, vol. 9, no. 4, e35882, 2022, doi: 10.2196/35882.
6. L. Laranjo et al., "Conversational agents in healthcare: A systematic review," J. Am. Med. Inform. Assoc., vol. 25, no. 9, pp. 1248–1258, 2018, doi: 10.1093/jamia/ocy072.
7. L. Tudor Car et al., "Conversational agents in health care: Scoping review and conceptual analysis," J. Med. Internet Res., vol. 22, no. 8, e17158, 2020, doi: 10.2196/17158.
8. M. Milne-Ives et al., "The effectiveness of artificial intelligence conversational agents in health care: Systematic review," J. Med. Internet Res., vol. 22, no. 10, e20346, 2020, doi: 10.2196/20346.
9. H. Bendotti et al., "Conversational artificial intelligence interventions to support smoking cessation: A systematic review and meta-analysis," Digit. Health, vol. 9, 2023, doi: 10.1177/20552076231211634.
10. Y. He et al., "Conversational agent interventions for mental health problems: Systematic review and meta-analysis of randomized controlled trials," J. Med. Internet Res., vol. 25, 2023, doi: 10.2196/43862.
11. H. Li, R. Zhang, Y.-C. Lee, R. E. Kraut, and D. C. Mohr, "Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being," NPJ Digit. Med., vol. 6, 2023, doi: 10.1038/s41746-023-00979-5.
12. S. Bedi et al., "Testing and evaluation of health care applications of large language models: A systematic review," JAMA, vol. 333, no. 4, pp. 319–328, 2025, doi: 10.1001/jama.2024.21700.
13. World Health Organization, Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva, Switzerland: WHO, 2021.