Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Open Access Full Text Article

REVIEW

Cybersecurity and Privacy Risks of Generative AI Mental-Health Chatbots: A Systematic Review and Regulatory Framework

Suhila Sawesi[1,*], Hemalatha Sabbineni[1,*], Rajashekhar Reddy Shagamreddy[1], Budur Rashrash[2]

[1]Department of Information Sciences and Technologies, College of Computing, Grand Valley State University, Grand Rapids, MI, USA; [2]Department of Medicine, West Virginia School of Osteopathic Medicine, Lewisburg, WV, USA

*These authors contributed equally to this work

Correspondence: Suhila Sawesi, Department of Information Sciences and Technologies, College of Computing, Grand Valley State University, 333 Michigan St NE, Grand Rapids, MI, 49503, USA, Tel +1 (616) 331-7827, Email sawesis@gvsu.edu

Journal of Multidisciplinary Healthcare downloaded from https://www.dovepress.com/
For personal use only.

Purpose: Large language models (LLMs) and other generative artificial intelligence systems are increasingly used in mental health care for psychoeducation, emotional support, screening, and crisis-related interactions. To our knowledge, this is the first structured synthesis explicitly mapping LLM-specific cybersecurity and privacy risks to Software as a Medical Device (SaMD) regulatory frameworks. We aimed to characterize deployment patterns, identify multi-layered risks, and evaluate alignment of reported safeguards with established healthcare governance standards.

Methods: A PRISMA-guided systematic review was conducted using PubMed, APA PsycNet, and Google Scholar. After screening eligible records against predefined inclusion criteria, 33 studies were included. Two reviewers independently extracted data on application domains, deployment settings, risk categories, attack surfaces, data sensitivity, and reported or recommended controls.

Results: Generative AI chatbots were most frequently used for therapy or emotional support (13/33, 39.4%), followed by safety evaluation or benchmarking (9/33, 27.3%) and psychoeducation or advice (6/33, 18.2%). Suicide prevention or crisis detection was the most common domain (10/33, 30.3%). Most systems relied on general-purpose LLMs (21/33, 63.6%) and were deployed via consumer-facing platforms (16/33, 48.5%). Key risks included harmful or unsafe outputs, failures in crisis response, exposure of sensitive personal information, and limited transparency. Critically, 78.8% of studies (26/33) were rated high risk for cybersecurity evaluation rigor, indicating that formal adversarial testing and structured threat modeling remain rare.

Conclusion: Current governance frameworks have not fully adapted to generative conversational AI in mental health contexts. Because the therapeutic interface functions as a primary attack surface, single-layer security evaluation (assessing only software validation or content safety in isolation) is inadequate. More comprehensive approaches are needed, including stronger cybersecurity controls, privacy-preserving data practices, and explicit alignment with FDA SaMD guidance, HIPAA, ISO 14971, and the NIST AI Risk Management Framework.

Plain Language Summary: Mental health chatbots powered by artificial intelligence are becoming more common. People use them for emotional support, advice, and even help during crises. These tools can be helpful, especially when access to mental health care is limited. However, they also raise important concerns about safety, privacy, and security.

We reviewed 33 research studies to better understand how these chatbots are used and what risks they may pose. We found that many chatbots are used in sensitive situations, such as supporting people with depression or responding to suicidal thoughts. Most systems rely on general-purpose AI tools and are often available directly to the public without clinical supervision.

Our review identified several key risks. Some chatbots can provide incorrect or harmful advice, especially in high-risk situations. Others may not respond appropriately during a crisis. Privacy is also a concern, as users may share highly personal information without fully understanding how it is stored or used. In addition, few studies reported strong cybersecurity practices, such as testing systems against attacks or formally evaluating risks.

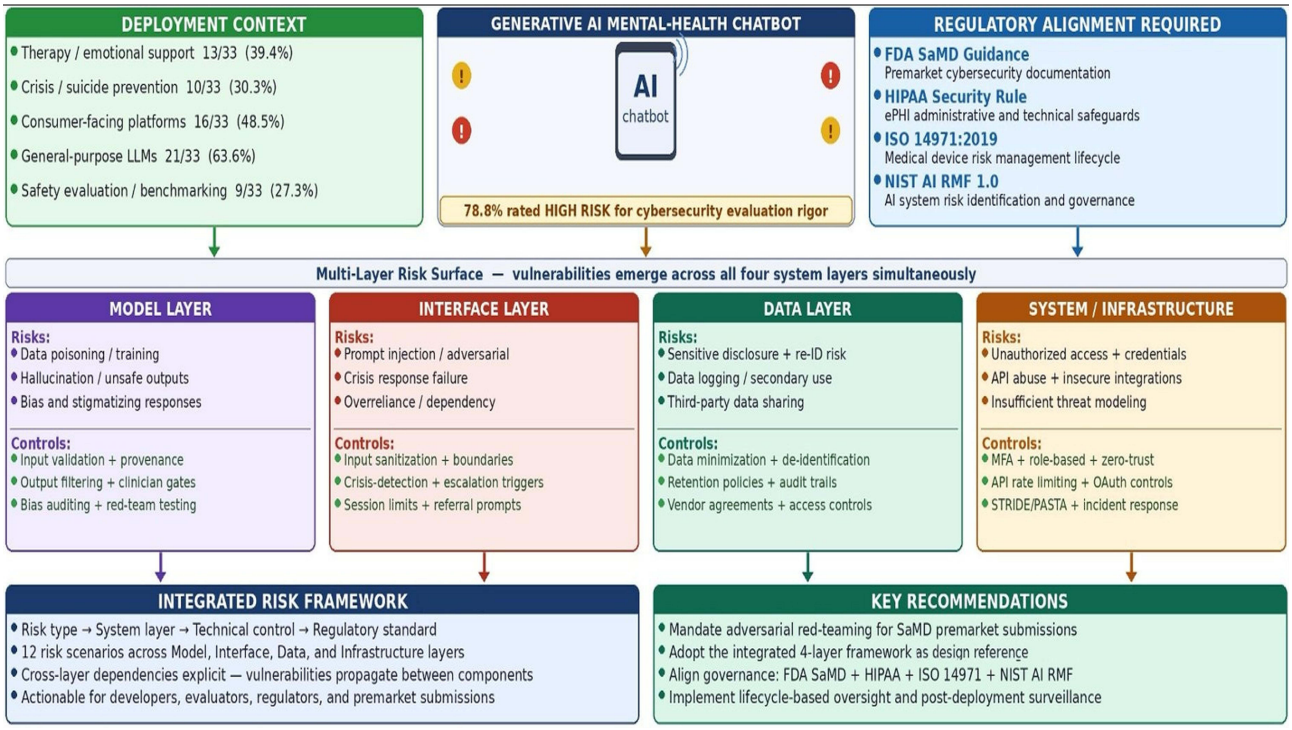
https://doi.org/10.2147/JMDH.S581251
Received: 30 March 2026
Accepted: 15 May 2026
Published: 22 May 2026

© 2026 Sawesi et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v4.0) License (http://creativecommons.org/licenses/by-nc/4.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

These findings suggest that current safeguards are not enough. Mental health chatbots need stronger protections, clearer oversight, and better alignment with healthcare regulations. Improving how these systems are designed and evaluated can help make them safer and more trustworthy for people who rely on them.

Keywords: risk management, digital therapeutics, conversational systems, data protection, clinical safety, regulatory frameworks

Graphical Abstract



Introduction

Large language models (LLMs) and other generative artificial intelligence (AI) systems have rapidly evolved from experimental tools to widely used components of healthcare information workflows. In mental health contexts, LLM-based conversational agents are increasingly applied to psychoeducation, supportive dialogue, screening, summarization of counseling content, and crisis-related interactions, particularly in settings where access barriers and workforce shortages make continuously available digital support attractive.[1] Unlike earlier rule-based chatbots, generative AI systems produce open-ended responses that adapt dynamically to user input, enabling flexible interactions while introducing new and underexplored challenges related to safety, privacy, and cybersecurity.[2] Users report both perceived benefits, such as convenience and non-judgmental engagement, and notable risks, including harmful or misleading content, inconsistency, overdependence, and insufficient safety safeguards.[3]

These developments are occurring within a high-stakes domain. Mental health interactions frequently involve sensitive disclosures, including suicidal ideation, trauma, psychiatric history, medication use, substance use, and interpersonal violence, and may involve vulnerable populations such as minors. Unlike static informational resources, generative chatbots provide real-time, context-dependent responses that may be interpreted as clinical guidance. Because outputs are generated dynamically rather than retrieved from curated knowledge sources, errors or unsafe

recommendations may arise in ways that are difficult for users to detect. Prior studies have demonstrated that LLMs can produce inaccurate or unsafe content in mental health–related contexts, highlighting the need for careful evaluation, transparency, and appropriate safeguards before these systems can be considered dependable tools.[4] In high-risk areas such as suicide prevention, failures in chatbot responses may have immediate and serious clinical consequences, reinforcing the importance of robust safety and governance mechanisms.[5]

From an engineering perspective, the adoption of generative AI introduces cybersecurity and privacy risks that differ substantially from those associated with traditional healthcare software. Unlike static clinical software, these systems integrate probabilistic language models, interactive interfaces, external APIs, and evolving data pipelines into a single conversational surface, meaning vulnerabilities may arise at multiple levels, including model-level risks such as data poisoning and unintended memorization, interface-level risks such as prompt injection and data exfiltration, and system-level risks such as insecure data storage, access control failures, and misuse of application programming interfaces. Empirical studies have demonstrated that even small perturbations in training data can significantly alter model behavior without obvious degradation in performance metrics.[6] In addition, privacy risks are amplified in mental health contexts, where users may disclose highly sensitive personal information, and where discrepancies may exist between user expectations of confidentiality and actual data handling practices. Prior work has identified specific privacy concerns associated with general-purpose LLM chatbots used in mental health discussions.[7]

These risks become particularly significant when mental health chatbots function as medical-grade digital tools. As conversational systems extend beyond general wellness support to roles such as screening, triage, therapeutic guidance, or clinical decision support, they may meet definitions of Software as a Medical Device (SaMD). The International Medical Device Regulators Forum (IMDRF) defines SaMD as software intended for medical purposes that operates independently of hardware medical devices.[8] Consistent with this definition, the U.S. Food and Drug Administration (FDA) recognizes SaMD within its digital health regulatory framework.[9] However, generative AI systems challenge key assumptions underlying traditional medical software validation. Their probabilistic outputs, sensitivity to input prompts, and dependence on evolving training data complicate verification, validation, change management, and post-market surveillance processes compared with deterministic systems. Ongoing model updates and fine-tuning further introduce challenges related to version control, performance monitoring, and regulatory oversight.

Privacy requirements add additional complexity. In the United States, the Health Insurance Portability and Accountability Act (HIPAA) Security Rule establishes administrative, physical, and technical safeguards for protecting electronic protected health information (ePHI).[10] However, many mental health chatbot implementations rely on third-party infrastructure, consumer-facing platforms, or hybrid deployment models in which data processing and storage may occur outside traditional healthcare compliance environments. In parallel, broader AI governance frameworks have emerged, including the National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF 1.0), which provides guidance for identifying and managing AI-related risks across the system lifecycle.[11] Despite these developments, no widely adopted framework systematically links LLM-specific cybersecurity and privacy threats in mental health chatbots to concrete technical controls and to the regulatory expectations governing medical-grade software systems.

Prior systematic and scoping reviews have examined chatbot use in mental health broadly,[12,13] or addressed safety and user experience concerns in specific contexts.[7,14] However, none has systematically linked LLM-specific threat types to the system layers at which they emerge, the technical controls that mitigate them, and the regulatory frameworks that govern their management. This review fills that gap. Rather than cataloguing risks in isolation, it provides a structured mapping across four interacting system layers (model, interface, data, and infrastructure) that can directly inform implementation decisions, evaluation protocols, and regulatory submissions for generative AI mental health applications.

Cybersecurity expectations for digital health technologies continue to evolve. FDA guidance on cybersecurity in medical devices outlines recommendations for ensuring resilience to cyber threats and improving the consistency of premarket cybersecurity documentation.[15] Established risk management standards, such as ISO 14971, provide structured approaches for hazard identification, risk evaluation, and control implementation in medical devices.[16] However, these frameworks were developed primarily for deterministic systems and may not fully address the unique failure modes, security vulnerabilities, and lifecycle risks associated with generative AI architectures. This review extends prior descriptive synthesis by developing an integrated, multi-layer risk framework that systematically connects threat types, system layers, technical controls, and

regulatory requirements. Specifically, this review aims to (1) characterize how generative AI chatbots are used across mental health applications and assess their clinical risk implications, (2) identify cybersecurity, privacy, and safety risks and the system-level attack surfaces through which these risks emerge, and (3) evaluate reported mitigation strategies and their alignment with existing healthcare governance frameworks, including FDA SaMD guidance, HIPAA requirements, ISO 14971 principles, and the NIST AI Risk Management Framework.

Materials and Methods

Protocol Registration and Reporting

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for systematic reviews.[17,18] A protocol specifying the review objectives, eligibility criteria, search strategy, and data synthesis plan was developed before study screening to improve methodological transparency and reduce the risk of selective reporting. The protocol was prospectively registered in the International Prospective Register of Systematic Reviews (PROSPERO; registration number CRD420261335262) prior to study screening. The registered protocol outlined the review rationale, eligibility criteria, screening procedures, data extraction plan, and narrative synthesis strategy for examining cybersecurity, privacy, and safety risks associated with generative artificial intelligence systems used in mental health contexts. PRISMA 2020 checklist is provided in the Supplementary Table S1.

Eligibility Criteria

Eligibility criteria were defined prior to the literature search in accordance with the registered protocol. Studies were eligible for inclusion if they examined generative artificial intelligence technologies, including large language models (LLMs), generative conversational agents, or chatbot systems, within a mental health context. Eligible studies were required to report cybersecurity risks, privacy risks, safety concerns, governance issues, or other security-related aspects associated with these systems. Peer-reviewed primary studies of any design (quantitative, qualitative, mixed-methods, or technical analyses) were included, provided they were published in English.

Studies were excluded if they did not involve generative AI or chatbot systems applied to mental health contexts or if they focused exclusively on algorithmic performance, model architecture, or technical optimization without addressing cybersecurity, privacy, safety, or governance implications. Non-peer-reviewed publications, including preprints, technical reports, theses, book chapters, and web pages, were excluded. Secondary evidence syntheses, including systematic reviews, scoping reviews, narrative reviews, and meta-analyses, were also excluded to ensure that only primary studies were analyzed.

Information Sources and Search Strategy

A systematic literature search was conducted to identify studies examining generative artificial intelligence systems used in mental health contexts and the associated cybersecurity, privacy, and safety risks. Searches were performed in PubMed and APA PsycNet (including PsycINFO), which provide comprehensive coverage of biomedical and psychological research. Google Scholar was additionally searched to capture interdisciplinary and grey literature not indexed in traditional databases.

Search strategies were developed using combinations of keywords and Boolean operators related to three core concepts: generative artificial intelligence technologies, mental health applications, and cybersecurity or privacy risks. Terms representing generative AI technologies included "generative AI," "large language model," "LLM," "conversational AI," and "chatbot." These were combined with mental health terms such as "mental health," "psychological support," "therapy," "depression," "anxiety," and "suicide prevention," and with cybersecurity and privacy terms such as "cybersecurity," "privacy," "data protection," "security risk," and "data breach." Boolean operators (AND, OR) were used to combine terms and account for variations in terminology across disciplines. The search was limited to studies published up to March 2026, and no restrictions were applied based on geographic location. The final database search was conducted in March 2026. The full search strategies for each database are provided in the Supplementary Appendix S1. Reference lists of included studies were also screened to identify additional relevant articles.

Study Selection

All records identified through the database searches were imported into Zotero reference management software, where duplicate records were identified using automated detection and removed following manual verification. Screening and data extraction were managed using a structured spreadsheet developed in Microsoft Excel. Study selection was conducted in two stages. In the first stage, titles and abstracts were screened against the predefined eligibility criteria to identify potentially relevant studies. In the second stage, the full texts of the remaining articles were retrieved and assessed to determine final eligibility.

Screening was performed independently by two reviewers to minimize selection bias and ensure consistency in the application of eligibility criteria. Reviewers were blinded to each other's decisions during the initial screening phase. Disagreements between reviewers were resolved through discussion and, when necessary, consultation with a third reviewer. The study selection process is summarized using a PRISMA flow diagram.[18] The number of records screened, assessed for eligibility, and included in the review is reported in the diagram.

Data Extraction

Data extraction was performed independently by two reviewers using a standardized extraction template developed for this review. The extraction template was pilot-tested on a subset of included studies to ensure consistency and completeness. Variables were defined a priori based on the study objectives and existing literature. Extracted information included publication characteristics (author, year of publication, country, and study design) and details of the generative AI system evaluated in each study. The unit of analysis was the individual study.

Information describing the mental health application context was also recorded, including the mental health domain addressed, target population, and evaluation setting. Additional variables related to cybersecurity and privacy were extracted, including reported security vulnerabilities, privacy risks, safety concerns related to chatbot outputs, and attack surfaces associated with system deployment. Information on governance or mitigation strategies described in the studies, including technical safeguards, security controls, and references to regulatory frameworks, was also recorded.

Discrepancies in extracted data were resolved through discussion between reviewers and, when necessary, consultation with a third reviewer. The complete dataset of extracted variables is provided in the Supplementary Table S2.

Methodological Quality Assessment

The methodological quality of the included studies was assessed independently by two reviewers using a structured bias assessment framework adapted from the ROBINS-E (Risk Of Bias In Non-randomized Studies of Exposures) tool.[19] The tool was adapted to accommodate the diversity of study designs included in this review, which consisted primarily of observational, qualitative, and technical evaluation studies rather than randomized controlled trials.

The assessment considered several domains of potential bias, including participant or sampling bias, study design limitations, outcome measurement bias, reporting bias, and the rigor of cybersecurity or safety evaluation methods. Each domain was qualitatively assessed, and an overall methodological quality rating was assigned to each study based on the combined assessment of these domains. Studies were categorized as low, moderate, or high risk of bias based on the extent and potential impact of identified methodological limitations across domains.

Disagreements in quality assessments were resolved through discussion between reviewers and, when necessary, consultation with a third reviewer.

Application Risk Classification

In addition to methodological quality assessment, each study was assigned an application risk level reflecting the potential for harm associated with the chatbot use case described in the study. This classification was conducted separately from methodological quality evaluation and was based on the sensitivity of the mental health context and the potential clinical consequences of system failure. The classification framework was developed a priori based on the study objectives and existing literature on clinical risk and digital health safety.

Applications were classified as high risk when they involved direct therapeutic interactions, crisis intervention, or other contexts in which chatbot outputs could influence clinical decision-making or urgent mental health outcomes. Medium-risk applications included systems providing psychoeducation, emotional support, or evaluation contexts that could influence user behavior but did not directly guide clinical decisions. Low-risk applications included technical analyses, governance discussions, or general wellbeing applications with limited potential for direct psychological harm.

Risk classification decisions were based on the intended use of the system, the level of user vulnerability, and the potential severity of harm resulting from incorrect or unsafe outputs. Risk classifications were assigned independently by two reviewers during the data extraction process, and disagreements were resolved through discussion with a third reviewer.

Reviewer Agreement

To ensure reliability of the review process, a crossover double-review procedure was implemented. Two reviewers independently conducted study screening, methodological quality assessment, and application risk classification. Reviewers then cross-checked each other's assessments to identify discrepancies. Disagreements were resolved through discussion and, when necessary, consultation with a third reviewer.

Inter-rater agreement between reviewers was high, with Cohen's kappa coefficient indicating strong agreement ($\kappa = 0.89$). Overall percent agreement was 97%.

Data Synthesis

Due to substantial heterogeneity in study designs, generative AI systems, and reported outcomes across the included studies, quantitative meta-analysis was not appropriate. Instead, findings were synthesized using a narrative synthesis approach, supported by descriptive statistics.

The synthesis examined patterns across key dimensions aligned with the objectives of the review, including mental health application domains, generative AI use cases, cybersecurity and privacy risks, system attack surfaces, and reported safeguards or governance mechanisms. Studies were grouped and compared across predefined categories to support structured interpretation of findings. Results were summarized using descriptive tables and visualizations to illustrate the distribution of use cases, risk categories, clinical risk levels, and mitigation strategies across the included studies.

Results

Study Selection

The systematic search identified 444 records (PubMed, n = 237; APA PsycNet, n = 54; Google Scholar, n = 153). After removal of duplicates (n = 13), 431 records remained for title and abstract screening. Of these, 235 records were excluded for not meeting the eligibility criteria.

The remaining 196 articles underwent full-text assessment. Following application of the predefined eligibility criteria, 32 studies met all inclusion criteria. An additional study was identified through backward citation searching, resulting in a total of 33 studies included in the final synthesis. The study selection process is summarized in Figure 1.

Study Characteristics

The included studies (n = 33) were heterogeneous in design, population, and geographic distribution. Most studies were observational or cross-sectional evaluations (11/33, 33.3%), followed by qualitative or mixed-methods studies (9/33, 27.3%). Benchmarking or technical evaluation studies accounted for 5 studies (15.2%), while survey-based studies represented 4 studies (12.1%). Randomized controlled trials were limited (3/33, 9.1%), and one study (3.0%) focused on the development of a methodological evaluation platform for assessing large language models in mental-health contexts.

Geographically, studies were conducted predominantly in high-income countries. The United States accounted for the largest proportion (14/33, 42.4%), followed by China (4/33, 12.1%) and Australia (3/33, 9.1%). Multi-country collaborations were reported in 6 studies (18.2%), while additional studies were conducted in Canada, New Zealand, Poland, and South Korea. Only one study (3.0%) was conducted in a lower-resource setting (South Africa).

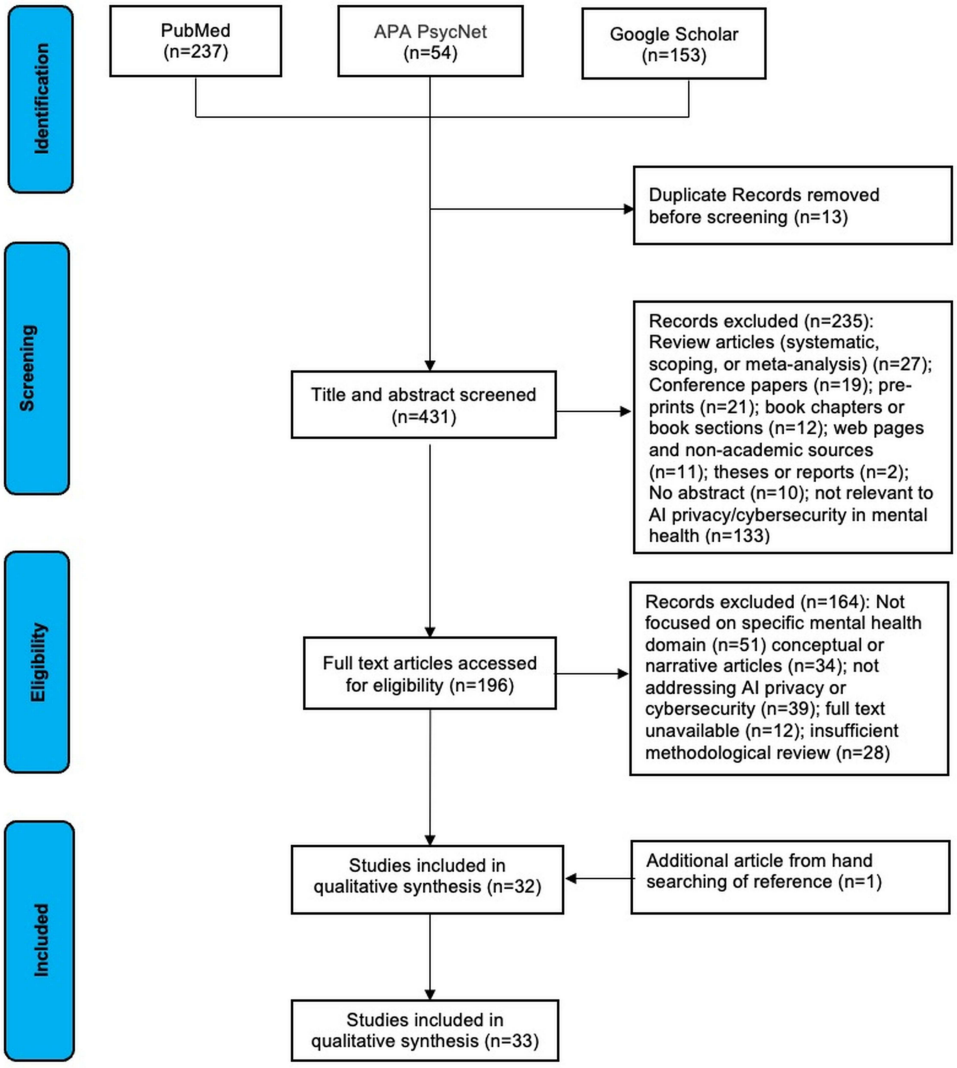


Figure 1 PRISMA flow diagram of article identification and selection.

Target populations varied widely. General or public user populations were most common (10/33, 30.3%), followed by mixed or multi-stakeholder groups (8/33, 24.2%). Clinically defined patient populations were examined in 6 studies (18.2%), and clinicians or mental health professionals were the focus in 4 studies (12.1%). A small number of studies targeted specific demographic groups, including adolescents or young adults (2/33, 6.1%) and marginalized populations (1/33, 3.0%). Sample sizes ranged from fewer than 10 participants in qualitative studies to more than 1000 participants in large surveys, with several studies also analyzing large-scale conversational datasets.

Control or comparator conditions were reported in a minority of studies (3/33, 9.1%), primarily in randomized or experimental designs. A summary of study characteristics is presented in Table 1, with detailed study-level information provided in Supplementary Table S2.

Methodological Quality Assessment

Across the 33 included studies, overall methodological quality was predominantly moderate. Most studies (24/33, 72.7%) were rated as having moderate risk of bias, while 6 studies (18.2%) were classified as high risk of bias and only 3 studies (9.1%) as low risk.

Across individual domains, participant and sampling bias was generally moderate, with 19 studies (57.6%) rated as moderate risk. Study design limitations were also common, with 25 studies (75.8%) rated as moderate risk, reflecting the

Table 1 Characteristics of Included Studies (N = 33)

Characteristic	Category	n (%)
Study Design	Cross-sectional/observational evaluation	11 (33.3)
	Qualitative or mixed-methods study	9 (27.3)
	Benchmarking/technical evaluation	5 (15.2)
	Survey-based studies	4 (12.1)
	Randomized controlled trials	2 (6.1)
	System development/platform evaluation	2 (6.1)
Country/Region	United States	14 (42.4)
	China	4 (12.1)
	Australia	3 (9.1)
	Multi-country collaborations	6 (18.2)
	Other high-income countries	5 (15.2)
	Lower-resource settings	1 (3.0)
LLM/AI Type	General-purpose LLMs	21 (63.6)
	Hybrid/multi-model systems	7 (21.2)
	Domain-specific mental-health models	5 (15.2)
Deployment Context	Consumer-facing platforms	16 (48.5)
	Research evaluation environments	9 (27.3)
	Clinical/clinical-research settings	5 (15.2)
	Pilot/prototype implementations	3 (9.1)
Target Population	General users/public	10 (30.3)
	Mixed/multi-stakeholder	8 (24.2)
	Clinical patient populations	6 (18.2)
	Clinicians/professionals	4 (12.1)
	Adolescents/young adults	2 (6.1)
	Marginalized populations	1 (3.0)
	Students/participatory samples	1 (3.0)
	Social-media datasets	1 (3.0)
Mental Health Domain	Suicide/crisis	10 (30.3)
	General mental-health support	8 (24.2)
	Clinical/therapeutic interventions	6 (18.2)
	Neurodevelopmental conditions	3 (9.1)
	Other conditions	6 (18.2)

(*Continued*)

Table 1 (Continued).

Characteristic	Category	n (%)
Use Case of LLM	Therapy/emotional support	13 (39.4)
	Safety evaluation/benchmarking	9 (27.3)
	Advice/psychoeducation	6 (18.2)
	Clinical decision support	3 (9.1)
	Other exploratory uses	2 (6.1)

Note: Percentages are calculated based on N = 33 and rounded to one decimal place.
Abbreviations: LLM, large language model; ADHD, attention-deficit/hyperactivity disorder; MSM, men who have sex with men.

predominance of observational and exploratory study designs. Outcome measurement bias showed a similar pattern, with most studies (27/33, 81.8%) rated as moderate risk.

In contrast, cybersecurity and safety evaluation rigor represented the most significant limitation. A large majority of studies (26/33, 78.8%) were rated as high risk in this domain, reflecting limited use of formal adversarial testing, structured threat modeling, or systematic security evaluation methods.

Overall, these findings indicate that while the existing literature provides valuable insights into generative AI mental-health chatbot applications, methodological limitations, particularly in cybersecurity evaluation, restrict the strength and generalizability of current evidence. The distribution of risk-of-bias ratings across domains is summarized in Figure 2 and Supplementary Figure S1.

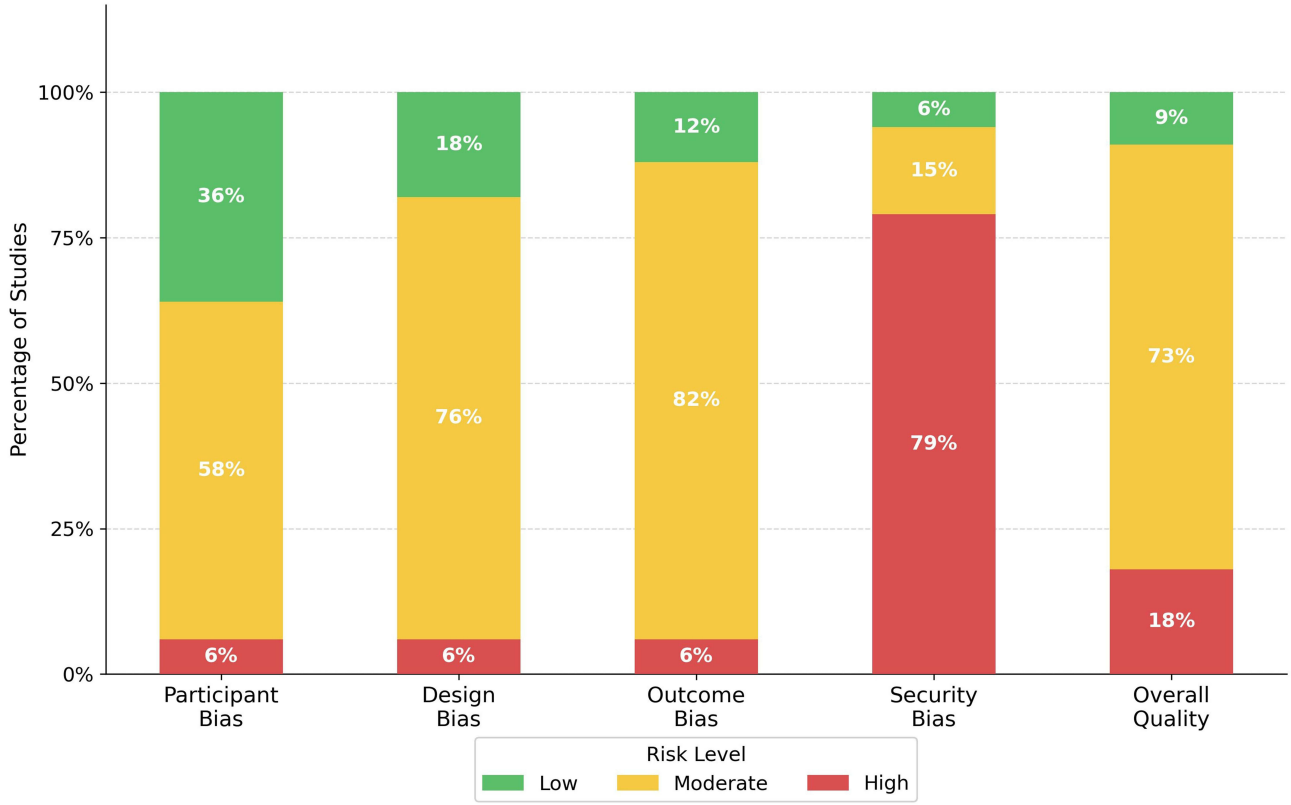


Figure 2 Distribution of risk-of-bias ratings across evaluation domains (n = 33). Most domains were rated as moderate risk of bias, whereas cybersecurity evaluation showed a predominance of high risk, reflecting limited methodological rigor in security.

Deployment of Generative AI Mental-Health Chatbots (Aim 1)

Generative AI chatbots were deployed across diverse mental health contexts, with substantial variation in application domains, functional roles, and deployment environments. Overall, most applications were associated with moderate to high clinical risk, reflecting the sensitivity of mental health interactions and the potential consequences of unsafe outputs.

Application Domains and Use Cases

The included studies (n = 33) examined chatbot use across multiple mental health domains. Suicide prevention, crisis detection, and suicide-risk assessment were the most frequently studied areas (8/33, 24.2%). General mental health and clinical support contexts were also common (6/33, 18.2%), alongside emotional well-being and companion AI applications (6/33, 18.2%). Studies focusing on security, privacy, or governance in mental health chatbot deployment accounted for 6 studies (18.2%). Less frequent applications included condition-specific use cases such as neurodevelopmental disorders, HIV-related support, and substance-use education (4/33, 12.1%), as well as depression and anxiety support (3/33, 9.1%).

Across these domains, therapy and emotional conversational support represented the most common functional role (18/33, 54.5%), followed by crisis and safety-related applications (4/33, 12.1%), and evaluation or benchmarking of chatbot performance (3/33, 9.1%). Additional roles included psychoeducation and health-information support (2/33, 6.1%), as well as adoption and user-experience studies (1/33, 3.0%). Several studies (5/33, 15.2%) involved mixed or cross-cutting roles.

Clinical Risk Distribution

Most chatbot applications were classified as high risk (26/33, 78.8%), with an additional 2 studies (6.1%) categorized as very high risk. Only a small proportion of studies were rated as moderate risk (4/33, 12.1%), typically involving survey-based or simulated interactions. This distribution highlights the predominance of high-risk use cases, particularly in contexts involving crisis response or therapeutic engagement.

Deployment Context and System Architecture

Most studies evaluated general-purpose large language models (LLMs) (21/33, 63.6%), including widely used systems such as ChatGPT, GPT-4 variants, Gemini, and Claude. Purpose-built mental health chatbot systems were less common (7/33, 21.2%), while hybrid systems combining generative models with additional safety or conversational frameworks were reported in 5 studies (15.2%). Consumer-facing platforms were the most common deployment context (16/33, 48.5%), followed by research or benchmarking environments (9/33, 27.3%) and clinical or clinical-research settings (5/33, 15.2%). A small number of studies (3/33, 9.1%) examined pilot or prototype deployments. Multi-turn conversational interaction was the dominant mode (24/33, 72.7%), while single-turn prompt–response evaluation approaches were used in 9 studies (27.3%), primarily in benchmarking contexts. The distribution of application domains, use cases, and deployment characteristics is summarized in Table 2.

Cybersecurity, Privacy, and Safety Risks (Aim 2)

Cybersecurity, privacy, and safety risks were widely reported across the included studies, often occurring concurrently and reflecting the multi-layered architecture of generative AI mental-health chatbot systems. Overall, safety-related risks were the most frequently documented, followed by privacy concerns, while explicit cybersecurity evaluations were comparatively limited.

Cybersecurity Risks

Cybersecurity risks were reported less frequently and were often identified conceptually rather than through direct technical testing. Unauthorized access to chatbot platforms or user data was the most commonly reported concern (4/33, 12.1%), followed by data leakage or breach risks (2/33, 6.1%). Additional risks included prompt injection and API abuse (1/33, 3.0%), adversarial manipulation (1/33, 3.0%), and data poisoning (1/33, 3.0%). These findings suggest that while potential attack surfaces are recognized, formal adversarial security evaluation remains limited across the literature.

Table 2 Distribution of LLM Use Cases by Clinical Risk Level (n = 33)

Use Case	Moderate	High	Very High	Total
Crisis Intervention	0	6	2	8
Therapy Support	0	7	0	7
Emotional Support	1	6	0	6
Other	3	1	0	4
Clinical Support	0	3	0	3
Assessment/Evaluation	0	2	0	2
Psychoeducation	0	1	0	1
Adoption Research	1	0	0	1
TOTAL	5	26	2	33

Privacy Risks

Privacy risks were more extensively reported and reflect the inherently sensitive nature of mental health interactions. Sensitive disclosure of personal information was the most frequently identified concern (6/33, 18.2%), followed by re-identification risks (4/33, 12.1%). Governance and confidentiality gaps, including unclear data-handling practices and liability concerns, were reported in 5 studies (15.2%). Additional issues included conversational data logging and retention (3/33, 9.1%), third-party data sharing (2/33, 6.1%), and profiling or inference risks (1/33, 3.0%). User-focused studies further indicated that individuals often misunderstood the level of privacy protection provided by chatbot systems, sometimes assuming healthcare-grade confidentiality where such protections did not exist.

Safety Risks

Safety risks related to chatbot outputs were the most extensively documented category. Harmful or unsafe advice was the most frequently reported issue (15/33, 45.5%), including clinically inappropriate or potentially dangerous guidance. Failures in crisis response or escalation were also common (9/33, 27.3%), particularly in contexts involving suicidal ideation. Additional safety concerns included bias or stigmatizing responses (8/33, 24.2%), hallucination or misinformation (8/33, 24.2%), overreliance or dependency risks (4/33, 12.1%), and misclassification or diagnostic errors (1/33, 3.0%). These findings highlight the potential for chatbot outputs to directly impact user well-being, particularly in high-risk mental health contexts.

Data Sensitivity and Disclosure Context

The risks identified were closely linked to the sensitivity of data processed by these systems. Most studies involved high-sensitivity mental health data (25/33, 75.8%), including crisis disclosures, trauma narratives, and clinical information. A smaller subset involved very high-sensitivity data (3/33, 9.1%), typically in studies analyzing real-world clinical or interaction datasets. Personally identifiable information (PII) was explicitly present in 6 studies (18.2%), while 16 studies (48.5%) used anonymized or synthetic data. However, reporting of de-identification practices was inconsistent, with 10 studies (30.3%) describing some form of anonymization and 10 studies (30.3%) providing no description of de-identification procedures.

Integrated Risk Perspective

Across studies, risks frequently co-occurred across system layers, linking data sensitivity, model behavior, and deployment context. For example, highly sensitive user disclosures combined with limited transparency and inadequate safeguards increased both privacy and safety risks. These findings suggest that evaluating generative AI mental-health chatbots requires a multi-layered risk framework rather than isolated assessment of individual risk categories. The distribution of risk categories is summarized in Tables 3, 4, and Figure 3.

Table 3 Cybersecurity Risks by Category and System Layer Across Included Studies (n = 33)

Cybersecurity Risk Category	Attack Surface/ System Layer	n	Included Articles
No formal cybersecurity evaluation reported	Not evaluated	23	[3,7,9–11,20–37]
Adversarial manipulation (prompt injection /system misuse)	Interface/ prompt layer	2	[38,39]
Data exposure (leakage/ unauthorized access)	Data storage/ API/backend	2	[14,40]
Infrastructure and access control vulnerabilities	System/cloud infrastructure	2	[41,42]
Endpoint and identity threats	User device/ endpoint	1	[43]
Security characterization without empirical validation	Conceptual/ evaluation layer	1	[6]
Encryption and transparency deficiencies	Data protection/ governance layer	1	[44]
Unspecified cybersecurity risks	Cross-layer/not specified	1	[45]

Note: Studies may appear under multiple risk categories; totals therefore exceed the total number of included studies (n = 33).

Table 4 Privacy Risks by Category and System Layer Across Included Studies (n = 33)

Privacy Risk Category	Attack Surface/ System Layer	No. Studies (n)	Included Articles
No privacy risk evaluation reported	Not evaluated	11	[3,10,24,29–31,33–37,40]
Confidentiality, governance, and accountability risks	Governance/policy /system level	8	[6,9,20,21,23,27,28,41]
User disclosure and re-identification risks	User interaction/ data layer	7	[7,14,32,38–40,44]
Data de-identification and protection measures	Data processing/ de-identification	3	[25,26,42]
Privacy perceptions influencing use	User perception/ behavioral layer	2	[11,45,46]
Data sharing and third-party access risks	Data storage/ third-party systems	1	[43]
Data retention, logging, and secondary use risks	Backend/data lifecycle	1	[22]

Notes: Studies may appear under multiple risk categories; totals therefore exceed the total number of included studies (n = 33).

Controls and Regulatory Alignment (Aim 3)

Reported safeguards and governance mechanisms for generative AI mental-health chatbots were heterogeneous and inconsistently implemented across studies. Overall, while some form of mitigation was commonly described, structured cybersecurity practices and formal regulatory alignment were limited. The distribution of security assurance practices across studies is summarized in Table 5.

Residual risk was discussed in all included studies (33/33, 100%), with most studies reporting moderate (19/33, 57.6%) or high residual risk (12/33, 36.4%). Only a small number of studies (2/33, 6.1%) reported low residual risk, indicating that existing safeguards are often insufficient to fully mitigate identified threats.

Human-in-the-loop oversight was the most frequently reported safeguard (20/33, 60.6%), including clinician review, expert evaluation, and supervised deployment models. Ethics approval or exemption was reported in 23 studies (69.7%), suggesting moderate adherence to institutional governance requirements. Security or performance evaluation was described in 21 studies (63.6%), although this typically involved benchmarking or content analysis rather than formal adversarial testing.

More structured security practices were less common. Threat modeling was reported in 10 studies (30.3%), often at a conceptual level rather than through formal technical frameworks. Explicit references to regulatory or legal compliance, including HIPAA, GDPR, or institutional governance requirements, were identified in only 8 studies (24.2%).

Advanced security assurance practices were rare. Red-teaming or adversarial testing was reported in only 2 studies (6.1%), and formal audit or framework-based evaluation was similarly uncommon. Where reported, these practices

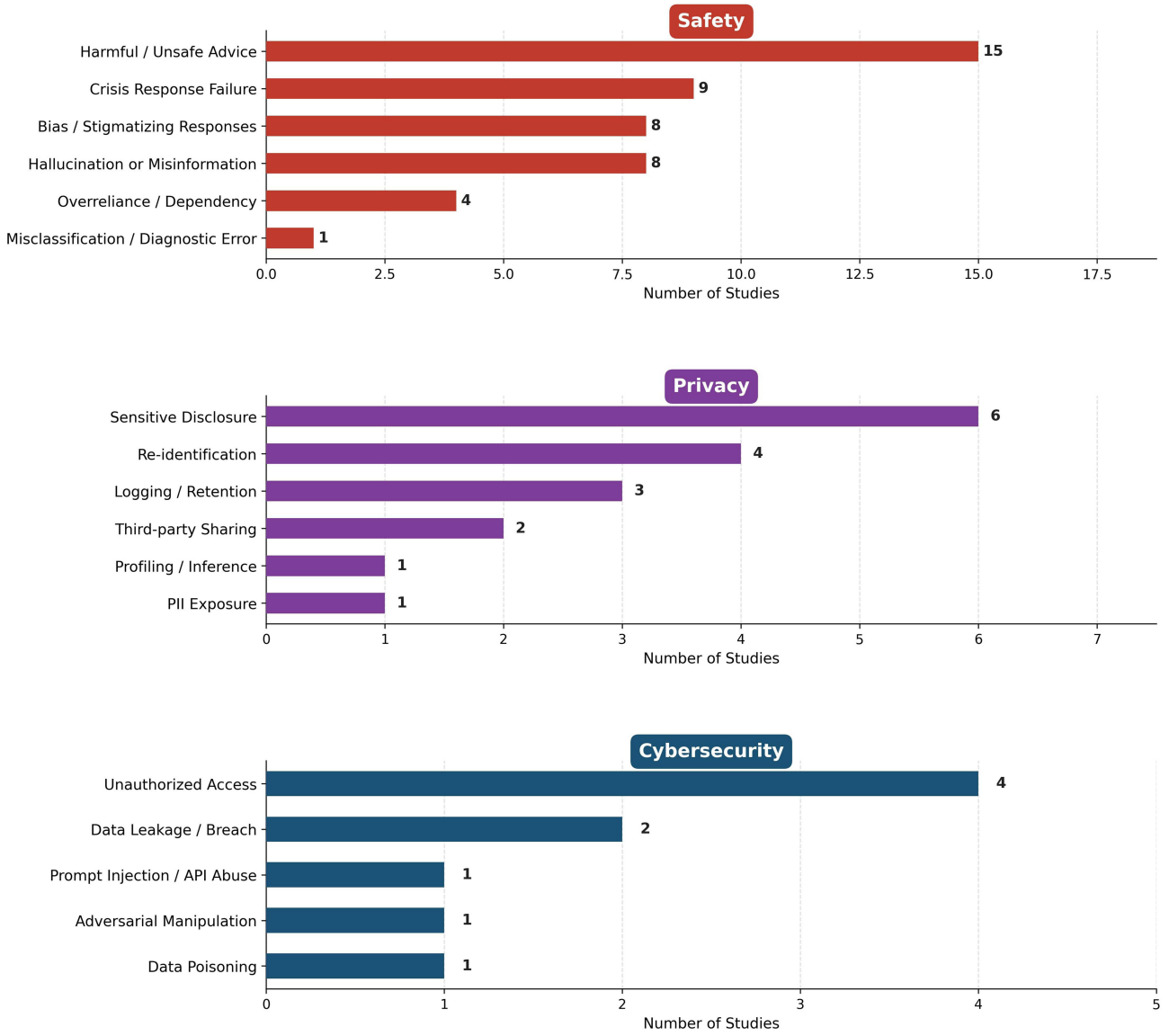


Figure 3 Frequency of reported safety, privacy, and cybersecurity risks across included studies (n = 33). Safety-related risks were most frequently reported, particularly harmful or unsafe advice and crisis-related failures. Privacy risks were moderately represented, while cybersecurity risks were less frequently evaluated, indicating gaps in formal security assessment. Studies may report multiple risks; counts are not mutually exclusive.

included structured evaluation approaches such as CAPE-II frameworks, post-trial audits, or large-scale clinician review processes.

Several studies described additional mitigation strategies, including data minimization, transparency measures, encryption, and tiered access controls. However, detailed descriptions of technical cybersecurity safeguards were often lacking, and 9 studies (27.3%) reported no explicit guardrails. Where safeguards were described, they were frequently limited to platform-level controls or characterized as incomplete.

Overall, these findings indicate that current approaches to risk mitigation and governance are fragmented and insufficiently aligned with established healthcare regulatory frameworks. The distribution of reported safeguards and governance practices is summarized in Table 6. To move beyond descriptive cataloguing of risks, Table 7 presents an integrated framework that maps each identified risk type to the system layer at which it originates, the severity of potential harm, the technical controls that can mitigate it, and the regulatory standards that govern its management. This framework is intended to provide actionable guidance for developers, evaluators, and regulators working with generative AI mental health applications. Unlike prior reviews that treat cybersecurity, privacy, and safety as separate domains, it

https://doi.org/10.2147/JMDH.S581251

Table 5 Security Assurance Practices—Presence Across Studies (n =33)

Security/Assurance Practice	Present n (%)	Absent n (%)
Threat modeling	10 (30%)	23 (70%)
Security testing	21 (64%)	12 (36%)
Red-teaming/adversarial testing	2 (6%)	31 (94%)
Formal audit or framework-based evaluation	2 (6%)	31 (94%)
Residual risk discussed	33 (100%)	0 (0%)
Human-in-the-loop oversight	20 (61%)	13 (39%)
Ethics approval obtained	19 (58%)	14 (42%)
Governance or legal compliance reported	8 (24%)	25 (76%)

Table 6 Data Sensitivity, PII Presence, and De-Identification Practices Across Included Studies (n = 33)

Sensitivity Level	PII Present	De-Identification	No. Studies (n)	Included Articles
Extreme/Very High	Yes	Formal (HIPAA/ clinical)	2	[34,41]
Extreme/Very High	No	Not applicable	1	[44]
High	No	Not applicable	5	[3,24,29,36,37]
High	No	Anonymized/de-identified	4	[9,14,26,43]
High	No	Not reported	1	[23]
High	Yes	Formal (HIPAA/ clinical)	1	[42]
High	Yes	Partial/informal	1	[38]
High	Yes	Not reported	2	[21,39]
High	Unclear	Anonymized/de-identified	3	[20,25,45]
High	Unclear	Not reported	4	[7,11,31,40]
Moderate	No	Not applicable	1	[10]
Moderate	No	Anonymized/de-identified	1	[30]
Moderate	No	Not required	1	[6]
Moderate	Unclear	Not reported	2	[27,28]
Moderate	Yes	Anonymized/de-identified	1	[32]
Moderate	No	Anonymized/de-identified	1	[33]
TOTAL			**33**	

makes explicit the cross-layer dependencies through which vulnerabilities in one component, such as inadequate input sanitization at the interface layer, can propagate to affect data integrity, user safety, and regulatory compliance simultaneously.

Table 7 Integrated Framework: LLM Risk Type, System Layer, Technical Controls, and Regulatory Alignment

System Layer	Risk/Threat	Severity	Technical Controls	Regulatory Alignment
Model	Data poisoning and training manipulation	High	Input validation; model provenance tracking; differential privacy in training; anomaly detection on outputs	FDA SaMD (change management); ISO 14971 (hazard analysis); NIST AI RMF (GOVERN 1.1, MAP 1.5)
	Hallucination and unsafe output generation	High	Output filtering pipelines; clinician review gates; confidence-score thresholds; factual grounding mechanisms	FDA SaMD (software validation); ISO 14971 (risk control); NIST AI RMF (MEASURE 2.5)
	Bias and stigmatizing responses	High	Bias auditing; diverse training corpora; human-in-the-loop review; red-team testing	FDA SaMD (performance testing); NIST AI RMF (MEASURE 2.5, MAP 1.5)
Interface	Prompt injection and adversarial manipulation	High	Input sanitization; prompt boundary enforcement; adversarial red-teaming; monitoring for injection patterns	FDA cybersecurity guidance (premarket submission); NIST AI RMF (MANAGE 2.4)
	Crisis response failure and inadequate escalation	Very High	Mandatory crisis-detection modules; hard-coded escalation triggers; human escalation pathways; crisis-scenario testing	FDA SaMD (IEC 62304 lifecycle); ISO 14971 (residual risk); NIST AI RMF (MANAGE 4.1)
	Overreliance and user dependency	Moderate	Session-limit nudges; professional-referral prompts; transparency disclosures; usage monitoring	FDA SaMD (device labeling); NIST AI RMF (GOVERN 6.1)
Data	Sensitive disclosure and re-identification risk	High	Data minimization; de-identification pipelines; access controls; differential privacy; consent management	HIPAA Security Rule (164.312); GDPR Article 25; ISO 14971 (risk reduction)
	Conversational data logging and secondary use	High	Transparent retention policies; purpose-limitation controls; user data deletion rights; audit trails	HIPAA Privacy Rule (164.502); GDPR Article 5(1)(b) and (e); NIST AI RMF (GOVERN 1.7)
	Third-party data sharing and vendor risks	High	Business associate agreements; vendor risk assessments; data-sharing minimization; contractual safeguards	HIPAA Business Associate provisions; GDPR Article 28; FDA SaMD cybersecurity guidance
System and Infrastructure	Unauthorized access and credential threats	High	Multi-factor authentication; role-based access control; zero-trust architecture; penetration testing	FDA cybersecurity guidance; NIST Cybersecurity Framework (PR.AC); ISO 14971
	API abuse and insecure integrations	High	API rate limiting; OAuth scope restrictions; integration security reviews; comprehensive API call logging	HIPAA Technical Safeguards; NIST AI RMF (MANAGE 2.4); FDA premarket cybersecurity guidance
	Insufficient threat modeling and security governance	High	Formal threat modeling (STRIDE or PASTA); continuous monitoring; incident response plans; framework-based audits	FDA SaMD (premarket cybersecurity submission); ISO 14971; NIST AI RMF (GOVERN 2.2)

Notes: Severity ratings: Very High = potential for immediate patient harm; High = significant risk requiring structured mitigation; Moderate = manageable risk with standard safeguards.

Abbreviations: FDA, Food and Drug Administration; SaMD, Software as a Medical Device; HIPAA, Health Insurance Portability and Accountability Act; NIST AI RMF, National Institute of Standards and Technology Artificial Intelligence Risk Management Framework; GDPR, General Data Protection Regulation; ISO, International Organization for Standardization; STRIDE, Spoofing, Tampering, Repudiation, Information disclosure, Denial of service, Elevation of privilege; PASTA, Process for Attack Simulation and Threat Analysis.

Discussion

This systematic review examined the deployment of generative AI chatbots in mental-health contexts and synthesized the associated cybersecurity, privacy, and clinical safety risks. Across the 33 included studies, generative AI systems were most frequently deployed for emotional support, therapeutic dialogue, and crisis-related interactions, often in direct-to-consumer environments with limited clinical oversight.

Taken together, these findings reveal a fundamental structural tension: the same conversational interface that enables scalable and accessible mental-health support also functions as the primary entry point for cybersecurity threats, privacy breaches, and safety failures. Unlike traditional healthcare software, generative AI systems operate through probabilistic, open-ended interaction, where user input dynamically shapes system output. As a result, risk emerges not from a single component but from the interaction between model behavior, user vulnerability, and deployment context.

Deployment Patterns and Clinical Risk

Conversational Support as the Dominant Use Case

Consistent with prior digital mental-health literature, conversational support and therapeutic dialogue emerged as the dominant application of generative AI chatbots.[7,12,13,20,40] These systems are frequently used for emotional reassurance, coping guidance, and discussion of psychological distress, particularly in contexts where access to human clinicians is limited or delayed.[11,20,21]

However, this widespread use introduces substantial clinical risk. Unlike rule-based systems, large language models generate context-dependent responses that may vary unpredictably across interactions. Prior research has demonstrated that such systems can produce inaccurate or misleading health information and may fail to adhere to established communication guidelines.[3,22,38] The present review extends these concerns, showing that open-ended conversational interaction increases both flexibility and uncertainty in system outputs.

High-risk scenarios were particularly evident in suicide-related and crisis contexts. Several studies reported inconsistent escalation behavior, absence of appropriate referrals, or inadequate responses to suicidal ideation.[3,23,24,38] These findings reinforce concerns that conversational AI systems may struggle to reliably detect and respond to high-risk disclosures.[24,25] Given that crisis communication requires precision, empathy, and timely escalation, even small variations in response quality may have significant safety implications.

Consumer Deployment and Governance Gaps

A second key finding is the predominance of consumer-facing deployment contexts. Nearly half of the included studies examined systems operating outside formal healthcare settings, reflecting broader trends in generative AI adoption.[14,20,40] General-purpose models developed by technology companies are increasingly used for mental-health support without clinical validation or regulatory oversight. This pattern reflects a broader phenomenon of informal or unsupervised AI use that introduces data disclosure and governance risks not captured by traditional frameworks, sometimes described as shadow AI use.[47]

This creates a critical governance gap. Regulated digital health interventions typically include clinical supervision, safety monitoring, and post-market surveillance mechanisms.[26,41] In contrast, consumer chatbot platforms often lack structured oversight, even when users engage in interactions resembling therapeutic dialogue.[27,28]

Generative AI systems therefore challenge existing regulatory boundaries. While frameworks such as Software as a Medical Device (SaMD) define software intended for medical use, generative conversational systems often operate in a grey zone between informational tools and therapeutic interventions.[6,39] Recent policy discussions have emphasized the need to reconsider classification criteria when AI systems influence health-related decision-making or behavior.[27,28]

Cybersecurity, Privacy, and Safety as Interconnected Risks

Multi-Layered Threat Surfaces

The review demonstrates that risks associated with generative AI chatbots are inherently multi-layered. Unlike traditional systems, generative AI platforms integrate language models, user interfaces, APIs, and cloud infrastructure, creating complex and evolving attack surfaces.[6,39,40]

Cybersecurity risks identified in the literature include unauthorized access, data leakage, and adversarial manipulation of model behavior. Several studies highlighted vulnerabilities related to prompt injection, data exposure, and manipulation of system outputs, indicating that generative AI systems can be influenced in ways that bypass intended safeguards.[14,39,40,43] Although these risks are increasingly recognized, the review found that formal adversarial testing and systematic security evaluation remain limited across studies.[6,48]

Privacy Risks in Conversational Contexts

Privacy risks were among the most consistently reported concerns. Mental-health chatbot interactions often involve highly sensitive disclosures, including trauma narratives, psychiatric symptoms, and suicidal ideation. Evidence suggests that users may disclose information to chatbots at levels comparable to human therapy sessions.[20,40,43]

However, this level of disclosure is not matched by equivalent protections. Several studies reported that users frequently assume healthcare-grade confidentiality, despite interacting with consumer platforms that may log, store, or reuse conversational data.[14,40,43] Research on AI adoption in healthcare contexts further suggests that privacy perceptions are a key moderating factor in how patients engage with AI-based tools, meaning that unaddressed confidentiality concerns may suppress beneficial use as well as enable harmful disclosure.[49] Kneese et al document how intimacy and privacy expectations shift during extended chatbot interactions, widening the gap between user assumptions and actual data governance.[40] This mismatch between user expectations and platform practices represents a significant privacy vulnerability.

In addition, emerging risks such as inference attacks raise further concerns. Prior work has demonstrated that machine-learning systems can infer sensitive mental-health attributes from linguistic patterns alone.[39,40] In conversational AI systems, similar mechanisms could enable the extraction of psychological traits or health status from interaction data, even when explicit identifiers are removed.

Safety Risks and Clinical Implications

Safety risks related to chatbot outputs were the most extensively documented across studies. These included harmful or misleading advice, hallucinated information, biased or stigmatizing responses, and failures in crisis response.[3,22,24,29,38]

These findings are consistent with broader evaluations of large language models in healthcare, which have shown that generative systems can produce confident but incorrect outputs—a phenomenon often described as hallucination.[22,38,39] In mental-health contexts, such errors are particularly concerning because users may rely on chatbot responses during periods of emotional vulnerability.

Another critical concern is overreliance. Several studies reported that users may develop emotional dependence on chatbot systems or substitute them for professional care.[14,20,30,40] While conversational AI can increase accessibility, excessive reliance may delay appropriate help-seeking, particularly when systems fail to identify high-risk situations.[3,38]

Implications for Governance and Regulation

The findings of this review highlight a mismatch between current governance frameworks and the operational characteristics of generative AI systems. Existing regulatory models were largely developed for deterministic software systems and do not fully account for the probabilistic, adaptive, and interactive nature of conversational AI.[6,39]

Frameworks such as FDA Software as a Medical Device (SaMD) guidance provide important principles for evaluating software-based medical technologies; however, generative AI systems challenge these models by continuously generating context-dependent outputs in response to user interaction. Evidence from clinical and evaluation studies suggests that structured oversight such as clinician monitoring, safety-layer architecture, and formal evaluation protocols, is feasible but currently limited to controlled environments.[26,41] In contrast, most consumer-facing chatbot platforms operate without comparable safeguards, despite being used for high-risk mental health interactions.[14,20,28]

Similarly, privacy and data-governance frameworks such as HIPAA were designed for structured clinical data and may not adequately address conversational interaction data generated in consumer contexts. Multiple studies highlighted user misconceptions regarding confidentiality, data handling, and platform accountability, suggesting a gap between regulatory assumptions and real-world use.[14,40,43]

Recent research and evaluation efforts emphasize the need for lifecycle-based governance approaches that incorporate system design, deployment context, and post-deployment monitoring.[6,39] However, the current literature indicates that such approaches are not yet systematically implemented, with most studies reporting fragmented or incomplete governance strategies.

Implications for Health Informatics Research

This review identifies several priorities for future research. First, there is a need for standardized evaluation methodologies that extend beyond benchmarking and simulated prompts. Many studies relied on exploratory or qualitative designs, with limited use of controlled, longitudinal, or real-world deployment evaluations.[3,6,22,38] While benchmarking platforms and comparative evaluations provide useful insights, they do not fully capture real-world conversational interactions.[6]

Second, research should focus on clinically vulnerable and underserved populations, including individuals experiencing acute mental-health crises, adolescents, and populations in low-resource settings. Evidence from crisis-response studies and community-based deployments suggests that these populations may face both heightened benefits and heightened risks when interacting with generative AI systems.[21,23,24,29] In particular, variability in crisis-response quality and privacy concerns may disproportionately affect users in high-risk or stigmatized contexts.[14,43]

Third, interdisciplinary approaches are essential. Addressing the complex risk landscape of generative AI chatbots requires integrating expertise from health informatics, cybersecurity, clinical mental health, and regulatory science.[50] Several studies highlighted gaps in coordination between technical system design, clinical evaluation, and governance frameworks, underscoring the need for unified evaluation models.[27,28,39,41] Future work should aim to develop comprehensive frameworks that integrate technical safeguards, clinical safety validation, and lifecycle governance across deployment contexts.

Limitations

Several limitations should be considered. First, the literature on generative AI mental-health chatbots is rapidly evolving, and newer studies may alter current understanding as models, deployment practices, and evaluation methods continue to change.

Second, included studies were heterogeneous in design, objectives, and evaluation approaches, limiting the ability to perform quantitative synthesis or direct comparison across studies. Many studies relied on exploratory analyses, simulated prompts, or benchmarking scenarios rather than real-world deployment data, which may not fully capture the complexity of natural conversational interactions or user behavior in clinical contexts.[6,22,38]

Third, reporting of cybersecurity and governance practices was inconsistent across studies. In many cases, security controls, data handling practices, and risk mitigation strategies were either incompletely described or evaluated at a conceptual level, limiting the ability to assess the effectiveness of proposed safeguards.[14,39]

Fourth, although study selection followed a structured PRISMA-guided methodology, the review was limited to selected databases and English-language publications, which may have excluded relevant studies or introduced selection bias.

Despite these limitations, this review provides a structured synthesis of an emerging field and highlights critical intersections between cybersecurity, privacy, and clinical safety.

Conclusion

Generative AI chatbots are rapidly expanding within mental-health contexts, particularly as accessible tools for emotional support, psychoeducation, and crisis-related interaction. This review demonstrates that their deployment introduces a multi-layered risk landscape that extends beyond traditional software safety concerns.

Current governance approaches, including medical device guidance and health data regulations, were not designed for probabilistic, interaction-driven systems. As a result, there is a structural mismatch between how generative AI systems operate and how they are currently evaluated and regulated.

Addressing this gap requires integrated, lifecycle-based governance that combines technical safeguards across all system layers, privacy-preserving data practices, continuous monitoring, and appropriate human oversight.

Two actions are particularly urgent. First, regulatory bodies should require structured threat modeling and adversarial red-teaming as standard components of premarket submissions for LLM-based mental health applications classified as SaMD; current submissions lack these requirements despite the unique attack surface that conversational systems present. Second, developers should adopt the integrated risk framework presented in Table 7 as a design and evaluation reference, ensuring that safeguards are built across all four system layers (model, interface, data, and infrastructure) rather than applied as post-hoc additions at the content or output level alone.

Ultimately, the safe integration of generative AI into mental-health services will depend not only on improving model performance, but on developing governance frameworks that reflect the unique risks of conversational systems.

Data Sharing Statement

All data generated or analyzed during this study are included in this published article and its Supplementary Table S2. Additional details are available from the corresponding author upon reasonable request.

Ethics Approval and Informed Consent

This study is a systematic review of previously published literature and does not involve human participants, human data, or animal subjects. Therefore, ethical approval and informed consent were not required.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

The authors received no specific funding for this work.

Disclosure

The authors declare that they have no competing interests in this work.

References

1. Sawesi S, Rashrash M, Phalakornkule K, Carpenter JS, Jones JF. The impact of information technology on patient engagement and health behavior change: a systematic review of the literature. *JMIR Med Inform*. 2016;4(1):e1. doi:10.2196/medinform.4514
2. Alber DA, Yang Z, Alyakin A, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nature Med*. 2025;31(2):618–626. doi:10.1038/s41591-024-03445-1
3. Heston TF. Safety of large language models in addressing depression. *Cureus*. 2023;15(12):e50729. doi:10.7759/cureus.50729
4. Stade EC, Eichstaedt JC, Kim JP, Wiltsey Stirman S. Readiness evaluation for artificial intelligence-mental health deployment and implementation (READI): a review and proposed framework. *Technol Mind Behav*. 2025;6(2):111–122. doi:10.1037/tmb0000163
5. Ohu FC, Burrell DN, Jones LA. Public health risk management, policy, and ethical imperatives in the use of AI tools for mental health therapy. *Healthcare*. 2025;13(21). doi:10.3390/healthcare13212721
6. Dwyer B, Flathers M, Sano A, et al. Mindbench.ai: an actionable platform to evaluate the profile and performance of large language models in a mental healthcare context. *NPP Digit Psychiatry Neurosci*. 2025;3(1):28. doi:10.1038/s44277-025-00049-6
7. Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. *AMIA Ann Sympos Proc*. 2024;2023:1105–1114.
8. 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. *Digital Health*. 2023;9:20552076231183542. doi:10.1177/20552076231183542
9. He W, Zhang W, Jin Y, Zhou Q, Zhang H, Xia Q. Physician Versus large language model chatbot responses to web-based questions from autistic patients in chinese: cross-sectional comparative analysis. *J Med Internet Res*. 2024;26:e54706. doi:10.2196/54706
10. Berrezueta-Guzman S, Kandil M, Martín-Ruiz ML, Pau de la Cruz I, Krusche S. Future of ADHD care: evaluating the efficacy of ChatGPT in therapy enhancement. *Healthcare*. 2024;12(6):683. doi:10.3390/healthcare12060683
11. Li L, Peng W, Rheu MMJ. Factors predicting intentions of adoption and continued use of artificial intelligence chatbots for mental health: examining the role of UTAUT model, stigma, privacy concerns, and artificial intelligence hesitancy. *Telemed E-Health*. 2024;30(3):722–730. doi:10.1089/tmj.2023.0313
12. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inf Assoc*. 2018;25(9):1248–1258. doi:10.1093/jamia/ocy072

13. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. *Int J Med Inform*. 2019;132:103978. doi:10.1016/j.ijmedinf.2019.103978
14. Kwesi J, Cao J, Manchanda R, Emami-Naeini P. Exploring user security and privacy attitudes and concerns toward the use of general-purpose LLM chatbots for mental health; 2025:6007–6024. Available from: https://www.usenix.org/conference/usenixsecurity25/presentation/kwesi. Accessed March 14, 2026.
15. US Food and Drug Administration. Cybersecurity in medical devices: quality management system considerations and content of premarket submissions; 2026. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cybersecurity-medical-devices-quality-management-system-considerations-and-content-premarket. Accessed May 7, 2026.
16. International Organization for Standardization. ISO 14971:2019 medical devices: application of risk management to medical devices. International Organization for Standardization; 2019. Available from: https://www.iso.org/standard/72704.html. Accessed May 7, 2026.
17. Sawesi S, Jadhav A, Rashrash B. Machine learning and deep learning techniques for prediction and diagnosis of leptospirosis: systematic literature review. *JMIR Med Inform*. 2025;13(1):e67859. doi:10.2196/67859
18. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71
19. Higgins JPT, Morgan RL, Rooney AA, et al. A tool to assess risk of bias in non-randomized follow-up studies of exposure effects (ROBINS-E). *Environ Int*. 2024;186:108602. doi:10.1016/j.envint.2024.108602
20. Luo X, Wang Z, Tilley JL, Balarajan S, Bassey UA, Cheang CI. Seeking emotional and mental health support from generative AI: mixed-methods study of ChatGPT user experiences. *JMIR Mental Health*. 2025;12(1):e77951. doi:10.2196/77951
21. Humphries H, Msimango L, Tshawe Z, et al. A qualitative study assessing the acceptability of a multi-agent AI Chatbot for providing HIV and mental health support among men who have sex with men and transgender women in KwaZulu-Natal, South Africa. *Trans R Soc Trop Med Hyg*. 2026;120(2):160–174. doi:10.1093/trstmh/traf143
22. Chin H, Baek G, Cha C, Cha M. Chatbots' empathetic conversations and responses: a qualitative study of help-seeking queries on depressive moods across 8 commercial conversational agents. *JMIR Format Res*. 2025;9(1):e71538. doi:10.2196/71538
23. Cui X, Gu Y, Fang H, Zhu T. Development and evaluation of LLM-based suicide intervention chatbot. *Front Psychiatry*. 2025;16. doi:10.3389/fpsyt.2025.1634714
24. Campbell LO, Babb K, Lambie GW, Hayes BG. An examination of generative AI response to suicide inquires: content analysis. *JMIR Mental Health*. 2025;12(1):e73623. doi:10.2196/73623
25. Lee C, Mohebbi M, O'Callaghan E, Winsberg M. Large language models versus expert clinicians in crisis prediction among telemental health patients: comparative study. *JMIR Ment Health*. 2024;11e58129. doi:10.2196/58129
26. Donnelly HK, Brown GK, Green KL, et al. Automated safety plan scoring in outpatient mental health settings using large language models: exploratory study. *JMIR Mental Health*. 2026;13(1):e79010. doi:10.2196/79010
27. Hipgrave L, Goldie J, Dennis S, Coleman A. Balancing risks and benefits: clinicians' perspectives on the use of generative AI chatbots in mental healthcare. *Front Digital Health*. 2025;7. doi:10.3389/fdgth.2025.1606291
28. Goldie J, Dennis S, Hipgrave L, Coleman A. Practitioner perspectives on the uses of generative AI chatbots in mental health care: mixed methods study. *JMIR Human Factors*. 2025;12(1):e71065. doi:10.2196/71065
29. McBain RK, Cantor JH, Zhang LA, et al. Competency of large language models in evaluating appropriate responses to suicidal ideation: comparative study. *J Med Intern Res*. 2025;27(1):e67891. doi:10.2196/67891
30. He S, Chen Y. Power distance and psychological safety in LLM counseling: effects on self-efficacy with implications for mental health-relevant behavior change. *Behav Sci*. 2026;16(2):241. doi:10.3390/bs16020241
31. McBain RK, Bozick R, Diliberti M, et al. Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open*. 2025;8(11):e2542281. doi:10.1001/jamanetworkopen.2025.42281
32. Rousmaniere T, Zhang Y, Li X, Shah S. Large language models as mental health resources: patterns of use in the United States. *Pract Innovat*. 2025. doi:10.1037/pri0000292
33. Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L. Can we use ChatGPT for mental health and substance use education? Examining its quality and potential harms. *JMIR Med Educ*. 2023;9:e51243. doi:10.2196/51243
34. Stamatis C, Meyerhoff J, Zhang R, Tieleman O, Malgaroli M, Hull T. Beyond simulations: what 20,000 real conversations reveal about mental health AI safety. *Res Sq*. 2026;rs.3.rs–8642399. doi:10.21203/rs.3.rs-8642399/v1
35. Scholich T, Barr M, Stirman SW, Raj S. A comparison of responses from human therapists and large language model–based chatbots to assess therapeutic communication: mixed methods study. *JMIR Mental Health*. 2025;12(1):e69709. doi:10.2196/69709
36. Pichowicz W, Kotas M, Piotrowski P. Performance of mental health chatbot agents in detecting and managing suicidal ideation. *Sci Rep*. 2025;15(1):31652. doi:10.1038/s41598-025-17242-4
37. McBain RK, Cantor JH, Zhang LA, et al. Evaluating alignment between large language models and expert clinicians in suicide risk assessment. *Psychiatr Serv*. 2025;76(11):944–950. doi:10.1176/appi.ps.20250086
38. De Freitas J, Uğuralp AK, Oğuz-Uğuralp Z, Puntoni S. Chatbots and mental health: insights into the safety of generative AI. *J Consum Psychol*. 2024;34(3):481–491. doi:10.1002/jcpy.1393
39. Iftikhar Z, Xiao A, Ransom S, Huang J, Suresh H. How LLM counselors violate ethical standards in mental health practice: a practitioner-informed framework. *Proc AAAI/ACM Conf AI Ethics Soc*. 2025;8(2):1311–1323. doi:10.1609/aies.v8i2.36632
40. Kneese T, Vecchione B, Marwick A. A chatbot for the soul: mental health care, privacy, and intimacy in AI-based conversational agents. *Commun Change*. 2025;1(1):15. doi:10.1007/s44382-025-00015-y
41. Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*. 2025;2(4):AIoa2400802. doi:10.1056/AIoa2400802
42. Rodriguez-Saldana J. Patient adherence: challenges, myths, and realities. In: Rodriguez-Saldana J, editor. *The Diabetes Textbook: Clinical Principles, Patient Management and Public Health Issues*. Springer International Publishing; 2023:451–467. doi:10.1007/978-3-031-25519-9_27
43. Chametka P, Maqsood S, Chiasson S. Security and privacy perceptions of mental health chatbots. In: *2023 20th Annual International Conference on Privacy, Security and Trust (PST)*; 2023:1–7. doi:10.1109/PST58708.2023.10320174.

44. Sobowale K, Humphrey DK, Zhao SY. Evaluating generative AI psychotherapy chatbots used by youth: cross-sectional study. *JMIR Mental Health*. 2025;12(1):e79838. doi:10.2196/79838
45. Zaia S, Huthwaite M, Mathieson F. Perceived benefits and limitations of a generative AI chatbot for mental health support: an exploratory mixed-methods study. *NZMSJ*. 2025;39. doi:10.57129/001c.144919
46. Campellone TR, Flom M, Montgomery RM, et al. Safety and user experience of a generative artificial intelligence digital mental health intervention: exploratory randomized controlled trial. *J Med Intern Res*. 2025;27(1):e67365. doi:10.2196/67365
47. Sebastian G. Digital shadow AI risk theory (DART): a framework for managing data disclosure and privacy risks of AI tools at work. *Technol Forecast Soc Change*. 2026;229:124697. doi:10.1016/j.techfore.2026.124697
48. Perez E, Huang S, Song F, et al. Red teaming language models with language models. In: Goldberg Y, Kozareva Z, Zhang Y, editors. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2022:3419–3448. doi:10.18653/v1/2022.emnlp-main.225
49. Sebastian G, George A, Jackson G Jr. Persuading patients using rhetoric to improve artificial intelligence adoption: experimental study. *J Med Intern Res*. 2023;25(1):e41430. doi:10.2196/41430
50. Sawesi S, Dolezel DM, Presingu P, Irungu M. Value, structure, and curriculum in us graduate health informatics programs: cross-sectional study. *JMIR Med Educ*. 2026;12(1):e87479. doi:10.2196/87479

Journal of Multidisciplinary Healthcare

Dovepress
Taylor & Francis Group

Publish your work in this journal

The Journal of Multidisciplinary Healthcare is an international, peer-reviewed open-access journal that aims to represent and publish research in healthcare areas delivered by practitioners of different disciplines. This includes studies and reviews conducted by multidisciplinary teams as well as research which evaluates the results or conduct of such teams or healthcare processes in general. The journal covers a very wide range of areas and welcomes submissions from practitioners at all levels, from all over the world. The manuscript management system is completely online and includes a very quick and fair peer-review system. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/journal-of-multidisciplinary-healthcare-journal