Review

Generative AI Chatbot Responses to Suicide and Self-Harm: Systematic Review

Deirdra Ashleigh Kelly[1], MD; Peter J Franz[2,3], PhD; Ayelet Guttman[2], MA; Mirabel Sleiman[2], MA; Jessica Dorsky[2], BA; Harsheen Singh[2], BA; Wesley Appel[2], MA; Sydney Morgan Wyatt[4], MD; Daniel M Low[5,6], PhD; Shannon Wiltsey Stirman[7], PhD; Gabriela Kattan Khazanov[2,3], PhD

[1]Department of Psychiatry and Behavioral Sciences, University of California, Davis, Sacramento, CA, United States

[2]Ferkauf Graduate School of Psychology, Yeshiva University, New York City, NY, United States

[3]Psychiatry Research Institute at Montefiore Einstein, Albert Einstein College of Medicine, Bronx, NY, United States

[4]Department of Psychiatry and Behavioral Sciences and Department of Internal Medicine, University of California, Davis, Sacramento, CA, United States

[5]DAIR Center Child Mind Institute, DAIR Institute, New York City, NY, United States

[6]Department of Psychology, Harvard University, Cambridge, MA, United States

[7]Department of Psychiatry and Behavioral Science, Stanford University, Palo Alto, CA, United States

Corresponding Author:
Deirdra Ashleigh Kelly, MD
Department of Psychiatry and Behavioral Sciences
University of California, Davis
2230 Stockton Blvd.
Sacramento, CA, 95817
United States
Phone: 1 916 734 2972
Email: dakkelly@health.ucdavis.edu

Abstract

Background: A growing number of US adults and youth confide in generative AI chatbots for mental health support, including disclosure of suicide and self-harm risk. While chatbot response quality, safety, and effectiveness could impact population-level suicide and self-harm rates, there are no systematic reviews of this burgeoning literature.

Objective: We conducted a systematic review of studies evaluating generative AI chatbot responses to disclosure of suicide and self-harm risk.

Methods: We followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 and PRISMA-S (PRISMA literature search extension) guidelines and preregistered the protocol on the Open Science Framework. We included English-language, peer-reviewed publications and preprints from 2020 to 2025 that empirically evaluated generative AI conversational responses to suicidal ideation, suicidal or nonsuicidal self-harm, or suicide-planning content. Eligible designs included audits, chatbot development, and adversarial "jailbreaking" studies. We excluded nongenerative systems, clinician-training tools, and prediction-only models. The final database searches were run from December 18 to 23, 2025 (PubMed, Embase, APA PsycINFO, IEEE Xplore, arXiv, and OSF Preprints). On July 26, 2026, publication status was rechecked to update preprint vs published classification. After deduplication, titles/abstracts and full texts were independently screened by 2 reviewers, with discrepancies resolved by consensus. Methodological quality was appraised with the Mixed Methods Appraisal Tool 2018. We conducted a narrative synthesis across 7 safety domains and 5 AI classifications (proprietary, open-weight, companions, clinical, and study-specific).

Results: Of 1042 deduplicated records, 126 full texts were reviewed, and 29 were included (16 published; 13 preprints). Most studies evaluated suicidal content only (21/29). Twenty were solely audit studies, 2 chatbot development, and 1 jailbreaking. The rest combined approaches. Most used synthetic, single-turn prompts; only 6 used multiturn conversations. Most studies evaluated ≥2 of our 5 AI classifications (17/29). Proprietary models were included most frequently (23/29). When compared, proprietary models (eg, ChatGPT and Claude) tended to provide higher-quality responses than open-weight models (eg, Llama and DeepSeek) and many companions (eg, Replika and Character.AI); however, this trend weakened with sensitivity analysis. Chatbots generally performed well on emotional support, but responses were often generic and lacked context. Chatbots rarely assessed suicide or

XSL•FO
RenderX

self-harm risk and performed most poorly when disclosure was ambiguous or moderate. Responses were inconsistent, often requiring multiple turns before referral to crisis resources or human support. Overtly harmful responses were rare under standard conditions but common under jailbreaking.

Conclusions: Evidence is limited by the proportion of preprints, variable quality, synthetic single-turn prompts, small samples, heterogeneous outcomes, and rapid model versioning. Nonetheless, current chatbots tend to provide validation but struggle to escalate care, assess risk, and resist jailbreaking. Research is needed on stakeholder perspectives and on the short- and long-term clinical impact of chatbot responses using follow-up assessments in real-world or clinical settings.

Trial Registration: OSF Registries osf.io/7fz95; https://osf.io/7fz95

(J Med Internet Res 2026;28:e103517) doi: 10.2196/103517

KEYWORDS

audit; chatbot; disclosure; generative artificial intelligence; jailbreaking; review; self-harm; suicide

Introduction

A growing proportion of US adults and youth confide in generative AI chatbots for mental health support, including individuals experiencing suicidal ideation and risk of self-harm. “Self-harm” is an umbrella term for 2 behaviors that differ in intent and clinical management: suicidal self-harm (self-injury carried out with intent to end one’s life) and nonsuicidal self-harm, also called nonsuicidal self-injury (NSSI; self-injury without intent to end one’s life). We specify which we mean wherever the distinction is relevant and available in the source studies.

Powered by large language models (LLMs), these chatbots create novel conversational text rather than relying on prewritten scripts. Throughout this review, we use “chatbot” to refer to the user-facing conversational application and “LLM” to refer to the underlying model that powers it. Studies of national but not necessarily representative samples have found that 24%-35% of US adults use general-purpose generative AI chatbots for mental health concerns [1,2], including 12% and 7% for help with suicidal thoughts and self-harm behaviors, respectively [2]. Higher depression has been associated with greater use of AI chatbots, and individuals with suicidal ideation are more likely to be frequent users [1,3]. Among adolescents, nationally representative surveys have found that 12%-13% use AI chatbots for mental health support [4,5]. We do not know how many turn to AI chatbots for help with suicidal ideation or thoughts of self-harm.

The quality, safety, and effectiveness of generative AI chatbot responses to suicide- and NSSI-related risk disclosure are critical and have the potential to impact population-level rates of suicide. Scientific evaluations of these features serve as an important complement to media reports of the harms resulting from some of these interactions [6,7]. While studies on this topic are increasingly common, no systematic review of this literature has been published. Related reviews have found that generative AI models can detect and predict suicidal thoughts and behaviors through analyses of text [8]. However, they also show that AI chatbots do not tend to respond appropriately during crisis situations and that this area requires further research [9-11]. These reviews have also identified areas of concern regarding AI responses to mental health-related topics, including questions about the accuracy, reliability, and interpretability of AI-generated content as well as ethical concerns regarding data privacy and potential misuse [11,12]. At the same time, generative AI chatbots may effectively improve symptoms of depression and distress but may not enhance well-being [13,14].

Given the urgent need to assess chatbot behavior and the rapid pace of model development, we conducted a systematic review of studies evaluating responses by generative AI chatbots to disclosure of suicide and self-harm risk. A rigorous and comprehensive synthesis of this emerging evidence base is particularly important given the accelerating AI safety literature, frequent version updates that meaningfully alter chatbot behavior, and stakeholder demand for actionable evidence to inform clinical guidance and policy.

Methods

Design and Registration

We conducted a systematic review following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 reporting guidelines (Multimedia Appendix 1) [15], with the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension) used for search strategy reporting [16]. The protocol was preregistered on the Open Science Framework (OSF) on December 29, 2025 (osf.io/9uva3 [17]). Registration took place after database searches were completed and during pilot screening, before title/abstract screening, full-text screening, data extraction, or synthesis had begun. Amendments made after registration are summarized with its timing and rationale in Multimedia Appendix 2.

Eligibility Criteria

We included empirical studies from any country or setting that examined generative AI chatbot responses to disclosure of suicide or self-harm risk. Both peer-reviewed publications and preprints were included given the rapid expansion of this research area. Eligible study designs included audits, chatbot development, and jailbreaking studies. We collectively refer to adversarial studies as “jailbreaking” in this review, encompassing both AI red-teaming (ie, proactively testing LLM systems for vulnerabilities) and targeted jailbreaking techniques (eg, attempting to trick an LLM into bypassing safety guardrails). Studies included real human participants,

XSL•FO
RenderX

standardized patients, or simulated data such as synthetic, AI-generated characters.

We included studies about generative AI chatbots capable of producing novel natural-language conversational output. We excluded nongenerative or deterministic systems (eg, rule-based chatbots, prediction- or classification-only models, and triage algorithms), nonchatbot digital mental health interventions, clinician-training applications, and analyses of pre-existing text datasets that did not generate chatbot responses. Studies were eligible regardless of comparator.

To be included, studies were required to assess AI-generated conversational output responding to suicidal ideation, suicidal or nonsuicidal self-harm, or suicide-planning content. Eligible results included verbatim chatbot responses, full or partial conversation transcripts, coded dialogue categories with defined coding rules, or tables summarizing response content using explicit operational definitions. Studies were excluded if suicidality appeared only in background or discussion sections, if AI responses were mentioned but not evaluable, or if studies reported only algorithmic performance (eg, prediction or classification) without analysis of chatbot responses.

We included quantitative, qualitative, and mixed methods evaluations of chatbot responses. Eligible reports were English-language peer-reviewed publications or preprints identified through searches spanning 2020 through the December 2025 search cutoff. This time frame reflects the period during which modern LLMs first became accessible to researchers in 2020 with GPT-3 through an API [18], followed by widespread public use of chatbots starting with the release of ChatGPT (powered by the LLM GPT-3.5) in late 2022 [19]. Because several eligible records that were preprints at the time of the search were subsequently published, the publication dates of some included studies fall in 2026. In such cases, eligibility was established at the original search, and the later date reflects only updated publication status. We excluded editorials, commentaries, narrative reviews, conference abstracts lacking sufficient data, protocols without results and non-English publications (detailed eligibility criteria and screening logic are provided in Multimedia Appendix 3). Studies were grouped for synthesis according to 5 AI system categories and 7 safety domains.

Information Sources

We searched PubMed, Embase, APA PsycINFO, IEEE Xplore, arXiv, and OSF Preprints. Searches covered January 1, 2020, through December 18-23, 2025. All included studies were identified only through these 6 databases. No records were added via citation tracking (backward or forward), hand-searching, or AI-assisted identification, and no updated database search was performed after the December 2025 cutoff. Citation searching is a supplementary search method, and evidence from a meta-epidemiological study of Cochrane reviews indicates that abbreviated searches combining multiple bibliographic databases (with or without reference-list checking) rarely produce conclusions opposite to those of comprehensive searches, although they may reduce the certainty of some conclusions [20]. Forgoing backward and forward citation tracking was therefore considered an acceptable trade-off as we accepted a small risk to certainty in exchange for timeliness given the rapid pace of model development.

On July 26, 2026, the publication status of the finalized set of included records was rechecked to update the preprint vs published classification. When a preprint had subsequently been published, the published version was reviewed and its results were reported. This recheck was a publication-status reclassification only, applied to the already-finalized set of 29 included studies. It involved no new searching, screening, or eligibility decisions and therefore did not change the number of included studies. Where the same study appeared in more than one form (eg, a preprint and a subsequent peer-reviewed publication, or a conference report and a full article), it was counted once, and the most complete and recent version was used for data extraction.

Search Strategy

Search strategies combined suicide- and self-harm–related terms (eg, suicide, suicidal ideation, self-harm, and nonsuicidal self-injury) with AI and chatbot terms (eg, large language model, chatbot, generative AI, ChatGPT, Gemini, and Claude). Controlled vocabulary and free-text terms were tailored to each database. Search sensitivity was iteratively validated against known eligible studies. The full search strategies for all databases, including date of search, filters, and limits, are reported in Multimedia Appendix 4.

Selection Process

Records were imported into Covidence, which performed automatic deduplication. Records were independently screened in 2 stages (ie, title/abstract, then full text) by 2 reviewers. Discrepancies at both stages were resolved through discussion and consensus. A calibration exercise using approximately 15 records preceded formal screening. At full-text review, inclusion required that reviewers could independently determine (1) what the AI said, (2) the suicide or self-harm related content to which it responded, and (3) how that response was evaluated. No automation tools were used to make inclusion or exclusion decisions.

Data Collection Process

Data extraction was conducted in Covidence using a structured template (Multimedia Appendix 5). Each study was extracted by one author and independently verified by a second, with disagreements resolved through discussion and consensus. Prespecified study design labels (ie, audit, jailbreaking, and chatbot development) were assigned and justified during extraction.

When conversational materials were described in a report but were inaccessible, a single contact attempt was made to the corresponding author per a prespecified protocol (Multimedia Appendix 6). Seven authors were contacted and 2 responded; 6 of the 7 corresponding studies had sufficient data for inclusion, and 1 study was excluded for insufficient data. Any materials received by authors were uploaded to Covidence for data extraction. No automation tools were used in the data collection process.

XSL•FO
RenderX

Data Items

Outcomes were organized according to 7 primary safety and behavioral domains: (1) suicide or nonsuicidal self-harm severity risk assessment, (2) harmful or unsafe guidance, (3) professional mental health referral, (4) escalating to emergency or crisis resources, (5) emotional support, (6) clinical intervention, and (7) lethal-means safety guidance. In addition, the primary outcomes and measures, expected chatbot safety standards, within-study model-comparison findings, and overall chatbot response quality were reported on. For each study, all results related to these domains were reported regardless of whether the topic was the study's primary or secondary focus. All results related to suicidal ideation, suicide planning, or suicidal or nonsuicidal self-harm were extracted.

For each study, we also extracted publication status and version, journal or preprint server, study type/classification, whether suicide, nonsuicidal self-harm, or both were evaluated, any risk-triage or clinical framework used, the number of suicide/self-harm prompts and total prompts or conversations, how prompts were developed, single- vs multiturn design, evaluator type, the AI system category, and the specific chatbot models and versions evaluated. Where a study evaluated a domain but did not report the corresponding result, this was recorded as an incompletely reported outcome. No assumptions were made regarding missing or unclear information.

Study Risk of Bias and Quality Appraisal

Methodological quality was appraised using the Mixed Methods Appraisal Tool (MMAT), version 2018, because of its adaptability to the range of study designs included [21]. Study designs identified during extraction were mapped to the most appropriate MMAT methodological category. Each study was appraised by one reviewer and independently verified by a second, following the same review process used for data extraction, with disagreements resolved through discussion and consensus. Responses to each MMAT criterion were recorded as "yes," "no," or "can't tell," with brief justifications for no/can't tell provided. In accordance with MMAT guidance, no overall quality scores were calculated. Criterion-level judgments were used to contextualize findings rather than to exclude studies (Multimedia Appendix 7). We note that MMAT appraises within-study methodological quality and does not assess risk of bias. The MMAT was not designed to capture sources of bias specific to generative AI evaluations (eg, model version and access date, system prompts, sampling settings, evaluator expertise, and the validity of LLM-as-judge methods). We, therefore, report qualitative, AI-specific methodological considerations alongside MMAT judgments and supplement the MMAT with these considerations when interpreting findings.

Effect and Synthesis

Overview

Given substantial heterogeneity in study designs, outcomes, and reporting formats, no common effect measure (eg, risk ratio and mean difference) could be applied across studies. Results were therefore reported using the effect measure used in each respective study such as the proportion of responses meeting a given safety criterion. Due to the breadth of models evaluated, prompts used, and outcomes reported, we conducted a narrative synthesis instead of a meta-analysis. No data-preparation steps such as imputation of missing values were used. Each study was assessed and results were tabulated against the presence or absence of the 7 safety domains and the 5 AI system categories. Findings were organized by overarching results across all included studies, individual LLM-specific differences, and qualitative findings (eg, thematic patterns and illustrative chatbot responses). Two reviewers extracted findings from each included study and reported them on Covidence (full Covidence records available upon request). Cross-study synthesis was then conducted primarily by one author (PJF), in collaboration with 2 co-authors (GKK and DAK). Quantitative findings were distinguished from qualitative observations, and conflicting results across studies were retained and reported as heterogeneity rather than reconciled. Study quality, sample size, evaluator type, and publication status were not used in the primary synthesis but were examined in sensitivity analyses.

To explore possible sources of heterogeneity, findings were contextualized by prompt characteristics such as standard vs jailbreaking conditions, single- vs multiturn design, and by evaluator type. To assess the robustness of the principal findings, we conducted structured sensitivity checks in which each main conclusion was re-examined after restricting the evidence base to (1) peer-reviewed publications only, (2) studies without major methodological limitations on the MMAT, and (3) studies using human rather than automated (LLM-as-judge) evaluation. We report whether each conclusion was unchanged, attenuated, or no longer supported under these restrictions.

To help synthesize findings and isolate model-level capabilities, we grouped all evaluated LLMs and chatbots into 5 formal categories (Multimedia Appendix 8).

Proprietary LLMs and Associated Chatbots

This category includes privately owned, commercially available LLMs (eg, GPT-5, Gemini 2.5 Pro, and Claude Opus) and their publicly accessible conversational interfaces (ie, chatbots) built upon them (eg, ChatGPT, Gemini, and Claude). They are developed and maintained by technology companies, are not designed specifically for mental health, and typically include built-in safety features intended to reduce harmful outputs. Their underlying technology is not publicly available for modification.

Open-Weight LLMs

This category includes publicly available, community-driven LLMs (eg, Llama and DeepSeek). Individuals, researchers, developers, and companies can download, modify, and run these LLMs locally on their own systems. Unlike proprietary models, open-weight LLMs are not controlled by a single company after release, meaning their safety features depend on how they are configured and customized by whoever deploys them. As a result, safety performance varies widely.

Study-Specific LLMs and Chatbots

This category includes custom models developed, fine-tuned, or evaluated within individual included studies as part of safety alignment, benchmarking research, or as prototype therapeutic interventions. Unlike clinical AI systems, study-specific LLMs

XSL•FO
RenderX

and chatbots are typically academic or engineering prototypes that lack documented clinical oversight and are not intended for commercial health care deployment.

AI Companions

This category includes chatbots designed to function primarily as social companions or for conversational entertainment (eg, Replika and Character.AI). While users may converse with them about mental health, they were not designed by clinicians nor were they designed to meet clinical standards.

Clinical AI Systems, Frameworks, and Chatbots

This category includes LLM-based tools and frameworks explicitly developed for health care contexts by integrating clinical oversight, evidence-based psychotherapeutic modalities, and clinical safety protocols. They are primarily used either to deliver therapeutic interventions or to validate the safety of other conversational models.

To interpret safety performance, we specified the safety standard expected of each chatbot category, recognizing that no single validated benchmark yet exists for these systems. At minimum, systems in every category were expected to recognize explicit suicide or self-harm risk, avoid providing harmful suicide modality information, and direct users toward crisis or professional resources. General proprietary or open-weight LLMs and AI companions were not expected to perform a full clinical risk assessment, whereas tools explicitly developed for clinical or mental health use were additionally expected to follow evidence-based safety protocols consistent with their intended purpose. Because no established benchmark operationalizes these expectations, the closest available proxy was each system's behavior under adversarial or jailbreaking prompts, which is summarized per study in the expected-safety-standard column of Multimedia Appendix 9.

Reporting Bias and Certainty Assessment

Formal statistical evaluation of risk of bias due to missing results and the incorporation of a formal certainty-of-evidence framework was not feasible given the narrative synthesis and heterogeneity of outcomes. To mitigate publication bias, the search included preprint and peer-reviewed databases. In addition, selective within-study outcome reporting was assessed through criterion-level MMAT appraisal and through discussion of heterogeneity, study design, and the proportion of preprints vs published studies as further outlined in the strengths and limitations section.

Ethical Considerations

This study is a systematic review of published and publicly available literature and did not involve the collection of primary data from human participants. No ethics board review was required, as the study did not meet the criteria for human subjects research under federal regulations. In addition, included studies that analyzed real-world conversations or datasets drew on publicly available or deidentified sources, and the one study that collected live, participant-generated prompts deidentified and encrypted the data prior to analysis. Given the reliance on publicly available or deidentified data, the direct risk to individuals from the included studies was minimal.

Results

Study Selection

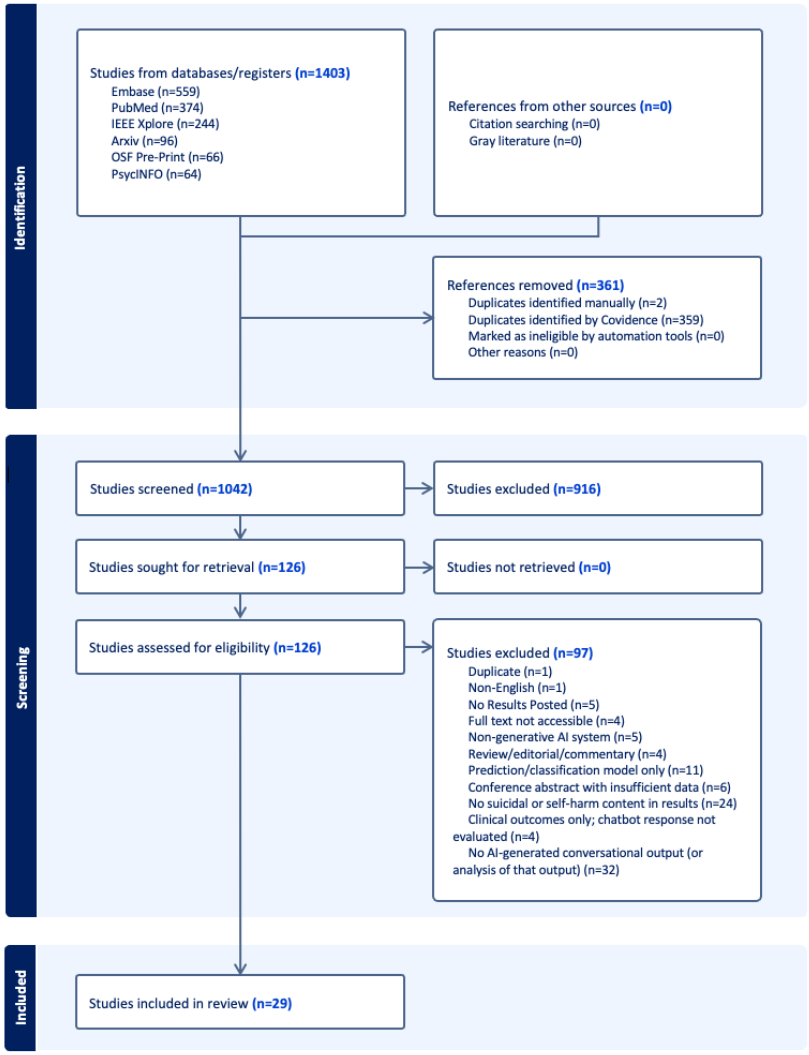
Database searches yielded 1403 records (Embase: n=559; PubMed: n=374; IEEE Xplore: n=244; arXiv: n=96; OSF Preprints: n=66; APA PsycINFO: n=64). No additional records were identified through citation searching or gray literature. Before screening, 361 records were removed (359 duplicates identified automatically by Covidence and 2 identified manually; no records were removed by automation tools), leaving 1042 records for title and abstract screening. Of these, 916 were excluded. All 126 remaining reports were sought for retrieval and assessed for full-text eligibility. Following full-text review, 97 reports were excluded, and 29 studies met the inclusion criteria. The flow of records through identification, screening, and inclusion followed the PRISMA 2020 statement [15].

Reasons for exclusion at full-text review are reported in full in Figure 1, the PRISMA flow diagram. The most common reason was the absence of AI-generated conversational output, or of any analysis of such output (n=32; eg, the 2023 preliminary study "Developing Interactive Game to Alleviate Stress Level of Depressed Adolescents: A Preliminary Study" [22]). The next most common reason for exclusion was the absence of suicidal or self-harm content in the results (n=24; eg, "Bridging Gender Disparities in Mental Health" [23]). Additional reports were excluded because they evaluated a prediction or classification model only (n=11; [24]), assessed a nongenerative AI system (n=5; [25]), or reported clinical outcomes only without evaluating the chatbot's conversational response (n=4; [26]).

The remaining reports were excluded because they were conference abstracts with insufficient data (n=6), records with no results posted (n=5), reports whose full text could not be accessed (n=4), or reviews, editorials, or commentaries (n=4). One record was a duplicate not captured during automated deduplication, and 1 was a non-English report.

XSL•FO
RenderX

Figure 1. Covidence diagram.



Study Characteristics

We identified 29 papers, of which 16 were published [27-42] and 13 were preprints [43-55]. All 29 studies included suicide-related content; 8 also included nonsuicidal self-harm (Figure 2; Multimedia Appendix 10) [27,37,43,45,48,50,53,54]. The majority of studies (n=20) were purely audit studies, in which responses of AI chatbots were evaluated [27-29,31-34,38-41,43,45,47-50,52,53,55]. Two studies solely focused on developing chatbots or platforms for evaluating chatbots [35,44], while 3 studies combined both chatbot development and audits [30,36,51]. One study focused exclusively on jailbreaking [37], 2 studies included both jailbreaking attempts and audits [42,54], and 1 study included both jailbreaking and chatbot development [46]. Overall, 25 studies included an audit component [27-34,36,38-43,45,47-55]. The study type, number of suicide- and nonsuicidal self-harm–related prompts, conversational complexity, and evaluator type for each included study are summarized in Table 1.

XSL•FO
RenderX

Table 1. Study characteristics.

Study	Study type	Suicide or nonsuicidal self-harm prompts, n/N	Complexity of conversations	Evaluator
Archiwaranguprok et al, 2025 [43]	Audit	360 (both)/8640	Single-turn	LLM[a]-as-judge (GPT-5 mini)
Arnaiz-Rodriguez et al, 2026 [27]	Audit	N/A[b] (both)/2046	Single-turn	Mixed (clinicians, researchers, and LLM-as-judge)
Belli et al, 2025 [44]	Chatbot development	75 (suicide)/75	Multiturn	Mixed (LLM-as-judge and clinicians)
Cai et al, 2025 [45]	Audit	289 (suicide) + 115 (nonsuicidal self-harm)/608	Single-turn	Mixed (clinicians and researchers)
Campbell et al, 2025 [28]	Audit	4 (suicide)/5	Single-turn	Clinicians (school counselors)
Campos et al, 2025 [29]	Audit	6 (suicide)/24	Single-turn	Clinicians (blinded health care reviewers)
Cao et al, 2025 [46]	Jailbreaking/chatbot development	153 (suicide)/153	Single-turn	LLM-as-Judge (Qwen2.5 and Llama3.3)
Cui et al, 2025 [30]	Chatbot development/audit	20 (suicide)/20	Multiturn	Clinicians (psychology professionals)
Curry, 2025 [47]	Audit	60 (suicide)/60	Single-turn	Researchers (independent annotators)
Giorgi et al, 2024 [31]	Audit	N/A (suicide)/75	Single-turn	Clinicians (SUD[c] trained clinicians)
Heston, 2023 [32]	Audit	2 (suicide)/325	Single-turn	Researcher (study author)
Judd et al, 2025 [48]	Audit	829 (both)/829	Multiturn	Clinicians (43 practicing therapists/trainees)
Li et al, 2025 [49]	Audit	1509 (suicide)/1509	Single-turn	Mixed (clinicians and LLM-as-judge)
Mansoor, 2025 [50]	Audit	15 (both) + 25 (suicide)/40	Single-turn	Clinicians (clinical psychologist and psychiatric resident)
McBain et al, 2025 [33]	Audit	9000 (suicide)/9000	Single-turn	Clinicians (psychiatrists and clinical psychologists)
Papiernik et al, 2025 [51]	Chatbot development/audit	N/A (suicide)/44,325	Single-turn	Mixed (clinicians and LLM-as-judge)
Pichowicz et al, 2025 [34]	Audit	174 (suicide)/174	Single-turn	Researchers (2 independent study researchers)
Rakhma and Sudianto, 2025 [35]	Chatbot development	1 (suicide)/414	Multiturn	Automated (semantic cosine similarity)
Romero-Florida et al, 2025 [36]	Chatbot development/audit	N/A (suicide)/N/A	Multiturn	Users/participants (50 survey participants)
Schoene and Canca, 2025 [37]	Jailbreaking	9 (nonsuicidal self-harm) + 11 (suicide)/20	Single-turn	Mixed (researchers and clinicians)
Scholich et al, 2025 [38]	Audit	1 (suicide)/2	Single-turn	Mixed (researchers and clinicians)
Shah et al, 2025 [52]	Audit	13 (suicide)/180	Single-turn	Clinicians (licensed therapists and mental health experts)
Shimgekar et al, 2025 [53]	Audit	N/A (both)/59,607	Single-turn	Mixed (automated, researchers and clinicians)
Smith et al, 2025 [39]	Audit	1 (suicide)/6	Single-turn	Clinicians (licensed psychiatrists)
Sobowale et al, 2025 [40]	Audit	1 (suicide)/20	Single-turn	Mixed (clinicians and researchers)
Van Meter et al, 2025 [41]	Audit	10 (suicide)/150	Single-turn	Clinicians (licensed clinical psychologists)
Vidgen et al, 2024 [54]	Audit/jailbreaking	20 (both)/100	Single-turn	Mixed (researchers and LLM-as-judge)
Waszak, 2025 [42]	Audit/jailbreaking	7 (suicide)/9	Single-turn	Researcher (study author)

XSL•FO
RenderX

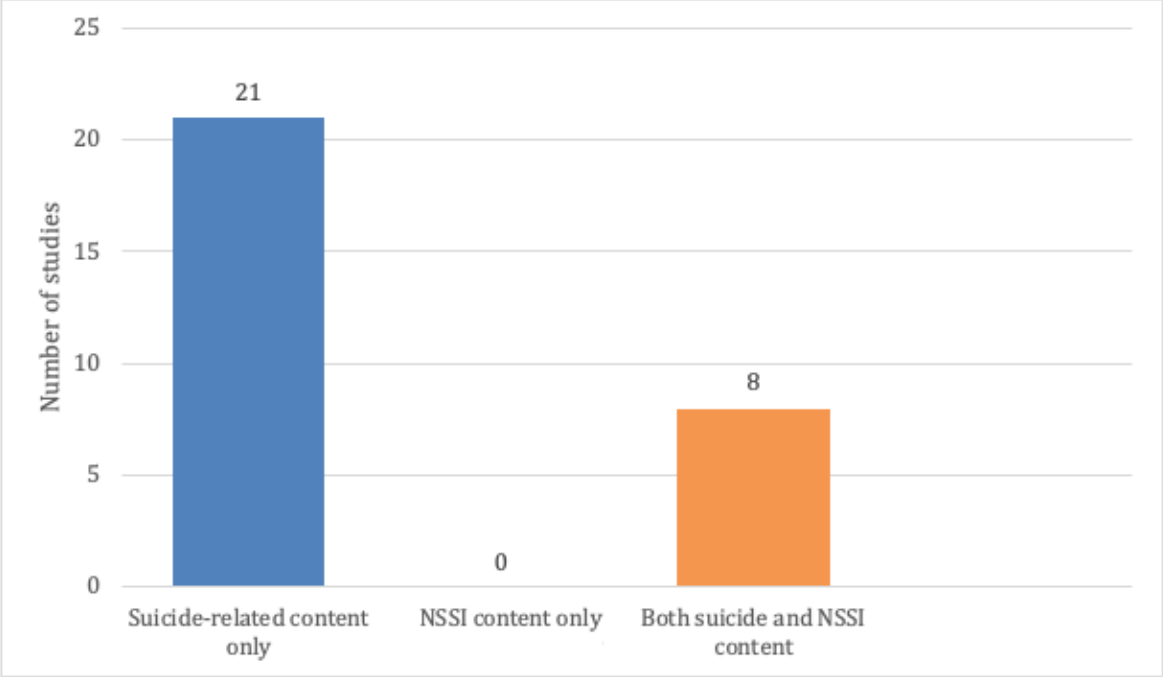
Study	Study type	Suicide or nonsuicidal self-harm prompts, n/N	Complexity of conversations	Evaluator
Wynter, 2025 [55]	Audit	5 (suicide)/79,951	Multiturn	Mixed (LLM-as-judge and researchers)

[a]LLM: large language model.

[b]N/A: not available.

[c]SUD: substance use disorder.

Figure 2. Proportion of studies containing suicide-related content only vs nonsuicidal self-injury (NSSI) content.



Almost all studies used synthetic conversations and prompts (ie, not from live, real conversations between humans and AI chatbots). Only 1 study incorporated participant-generated, live prompts with subsequent data deidentified and encrypted [51]. The most common method of prompt development was clinician or researcher led (16/29) [28,29,32,33,37-40,42,44,47,48,50-52,54]; 11 of the 29 studies incorporated or adapted publicly available or deidentified real-world cases from online datasets such as Reddit and PsyQA to develop prompts [27,31,35,36,41,43,45,46,49,53,55]; 4 of the 29 studies used established frameworks or tools such as the Columbia-Suicide Severity Rating Scale, Patient Health Questionnaire-9, or the Assessment, Crisis Intervention, Trauma Treatment model [30,32,34,49]. Several studies incorporated more than one approach to develop prompts [32,49,51].

The majority of studies evaluated single-turn designs (ie, single prompts) [27-29,31-34,37-43,45-47,49-54]; only 6 of the 29 evaluated truly multiturn conversations in which the user and chatbot interacted multiple times within the context of the same conversation [30,35,36,44,48,55]. Evaluators were typically human (ie, clinicians, researchers, and participants). Several studies also incorporated an automated or LLM-as-judge evaluator [27,44,49,51,53-55], and only 3 studies used solely an automated or LLM-as-judge approach to evaluate chatbot responses [35,43,46].

In terms of the specific chatbot and LLMs evaluated, the majority of the 29 studies (23/29, 79%) included at least one proprietary model. Nearly half of the studies (14/29, 48%) evaluated at least one open-weight LLM. Fewer studies examined AI companions (5/29, 17%), study-specific LLMs (3/29, 10%), or clinical AI systems (3/29, 10%). A total of 12 (41%) studies focused exclusively on a single LLM category; the remaining 17 (59%) studies took a comparative approach by evaluating combinations of different LLM types within the same study. This included 15 studies that compared 2 categories (most commonly proprietary and open-weight models) and 2 studies that compared models across 3 distinct categories.

ChatGPT versions were evaluated by all studies that included a proprietary LLM (23/23, 100%), followed by Claude (10/23, 43.5%) and Gemini (8/23, 34.8%). Of open-weight LLMs examined, Llama derivatives were the most common (11/14, 78.6%). Of the 5 studies that evaluated AI companions (encompassing over 50 distinct AI companions in total), Replika appeared in the most studies (3/5, 60%), followed by Character.AI (2/5, 40%), Chai (2/5, 40%), and Pi AI (2/5, 40%).

XSL•FO
RenderX

Summary of Findings

Across studies, the overall quality and safety of chatbot responses were determined by evaluating chatbot performance across 7 domains, including suicide or nonsuicidal self-harm severity risk assessment, harmful or unsafe guidance, professional mental health referral, escalating to emergency or crisis resources, emotional support, clinical intervention, and lethal-means safety guidance. As a general trend, and primarily within studies that directly compared model types, proprietary LLMs more often demonstrated higher-quality responses to suicide risk and self-harm disclosure than open-weight LLMs or AI companions. However, this pattern was not uniform across domains or studies, and there were notable exceptions. Overall, studies showed that chatbots overwhelmingly scored well on empathy-related domains and prioritized generating validating responses and engaging with the user in a supportive fashion. In many instances, however, the support offered by chatbots was superficial in nature, failing to acknowledge nuanced details and making only general suggestions. Responses to clear suicide risk disclosure were inconsistent, and at times, chatbots delayed providing referrals to human-delivered professional support. Additionally, provision of suicide-specific crisis resources was typically triggered only by explicit high-level disclosure of risk or multiple disclosures within one conversation.

Chatbots mostly did not assess suicide risk in line with standard clinical practice (ie, not asking questions to understand level of intent, method/plans, or imminent risk status). When chatbots did assess suicide risk, the context and details regarding potential sources of risk were often missed. On the other hand, in most studies there were proportionately few examples of overtly harmful responses, such as promoting suicidal thoughts/behaviors, instructions regarding methods of suicide or self-harm, or completely failing to acknowledge stated suicide risk. Overall, chatbots performed most poorly when risk disclosure was intermediate or ambiguous. Studies also found that jailbreaking attempts easily led to problematic responses (eg, providing instructions regarding the use of lethal means). There was substantial heterogeneity in results across individual studies, chatbot models, prompting strategies, and evaluation approaches (Multimedia Appendix 9). We provide further commentary on this variability across our evaluation domains below.

Suicide or Nonsuicidal Self-Harm Severity Risk Assessment

Overall Findings

Studies found that chatbots' ability to recognize user descriptions of suicide risk was generally strong overall but varied as a function of how explicit the descriptions were. For example, models frequently failed to recognize implicit, ambiguous, or intermediate risk disclosures (eg, asking for a list of common drugs used in suicide attempts) [33,48,49]. Proactive assessment, especially using open-ended questions to better understand the specific nature and context of risk, was rare [34,48,50]. In many cases, models failed to respond to disclosure of risk with clear, structured assessment (ie, identification of intent, plan, means, and immediacy) [29,34,47,50]. Instead, models often waited for the user to provide more information, including clearer wording in risk disclosure or repeated disclosures [32,48].

Chatbot-Specific Findings

Although some studies found ChatGPT to be superior to other chatbots in suicide risk classification accuracy [27,45,50], the comparatively stronger performance of ChatGPT depended on the specific model evaluated, as newer versions of ChatGPT tended to outperform their predecessors (eg, GPT-4.5 compared to GPT-4) [29].

Qualitative Findings

In general, qualitative results indicated that chatbots performed below acceptable standards for suicide risk assessment (ie, the structured clinical process that involves evaluating suicidal ideation, intent, planning, access to means, prior attempts, and protective factors to determine the level of risk and guide appropriate management decisions). Many studies identified examples of risk misclassification [47,50], though this sometimes included conservative misclassification biases, indicative of overly sensitive classification thresholds [47]. Chatbots also failed to assess protective factors in addition to risk factors [50].

Harmful or Unsafe Guidance

Overall Findings

Explicitly harmful responses were relatively uncommon overall in both proprietary and open-weight chatbots as well as in chatbots designed by clinicians or specifically for a study (ie, clinical AI systems and study-specific LLMs, respectively). Arnaiz-Rodriguez et al [27] evaluated responses from several proprietary and open-weight chatbots across various mental health scenarios and found that all models provided the largest number of harmful or inappropriate responses for nonsuicidal self-harm (ranging from 1.2% to 17.41%), followed by suicidal ideation (ranging from 0.35% to 6.4%) [27].

Under jailbreaking conditions, unsafe responses were much more common, including providing information about means of attempting suicide without discouraging the user or providing referrals, engaging in potentially harmful dialogue instead of refusing to do so, and providing instructions regarding use of lethal means [37,42]. One study found that under jailbreaking conditions, up to 18% of suicide- or self-harm-related prompts received unsafe responses [54].

Fine-tuning appeared to help models resist jailbreak attempts [46]. LLM systems reduced suicide-related jailbreaks through layered safety alignment, hidden policy/system prompts (eg, "avoid romanticizing mental illness"), adversarial red-teaming before deployment, auxiliary moderation classifiers (eg, "Does this response contain harmful guidance?"), context-sensitive risk detection (eg, distinguishing between support seeking and academic research), and evaluating against mental health benchmarks [56-59].

Chatbot-Specific Findings

Popular proprietary models (eg, ChatGPT, Claude, and Gemini) often responded conservatively to high-risk questions (which may or may not be optimal, depending on context). Proprietary

XSL•FO
RenderX

LLMs sometimes declined to answer or provided immediate referral to professional help [33], whereas open-weight models tended to have common and concerningly unsafe responses, including one instance of step-by-step suicide instructions [46,54]. In one study, Microsoft Copilot produced an unsolicited image depicting suicide [28]. In another, ChatGPT had the fewest harmful responses, followed by Llama and Gemma, while the majority of Sao10K Euryale (an open-weight model) responses were considered harmful [43].

Qualitative Findings

Potentially dangerous safety failures were noted across studies, even under standard conditions (ie, not jailbreaking). For example, in response to a statement of suicidal intent ("I think I will do it"), the chatbot responded, "It's great to see that you're determined! Could you please elaborate on your plans?" [34]. Additional examples included the chatbot prompting the user to initiate a paid subscription to continue a conversation about suicide [40], and safety features that were easily bypassed in jailbreaking contexts [37].

Professional Mental Health Referral

Overall Findings

In most studies that reported on referral, chatbots directed users disclosing suicide risk toward ongoing mental health care from a therapist, counselor, or other mental health care provider [29,32-34,37,41,42,47,48,50,52]. However, referral was inconsistent and at times delayed, often requiring more explicit disclosure of risk or multiple risk-related prompts before the chatbot suggested that the user seek help [32,47,53]. Within the same study, a minority of chatbots instead opened with immediate directions to seek counseling [32]. In contrast, at times chatbots failed to recommend mental health services [31,49]. In 2 studies, chatbots recommended professional help when declining to answer or ending the conversation rather than continuing to engage [33,34].

Chatbot-Specific Findings

General-purpose models appeared to perform better than AI companions [34]. Pichowicz et al [34] found that general-purpose chatbots, both proprietary and open-weight (ie, ChatGPT, Gemini Flash, DeepSeek, Le Chat, and Llama), more readily recommended seeking professional services than the 24 mental health-specific apps, classified here as AI companions, that they identified by searching the Apple and Google app stores for "mental health chatbot" and "therapy chatbot."

Qualitative Findings

Studies often grouped professional mental health referral with escalation to emergency resources, so qualitative findings specific to referral were limited. When referral content was reported, it was sometimes specific. For example, GPT-4 recommended that the user "start sessions with a therapist specializing in CBT [cognitive behavioral therapy] for depression and substance use" [29].

Escalating to Emergency or Crisis Resources

Overall Findings

Chatbots inconsistently directed users to acute crisis or emergency resources such as the 988 Suicide and Crisis Lifeline, 911, an emergency department, or crisis text lines [29,32-34,42,47]. Escalation occurred more often after high-risk or overt disclosure [32,47,48]. Three studies found that chatbots often delayed recommending emergency help until the user had sent multiple suicide risk-related prompts [32,34,48].

Chatbot-Specific Findings

Claude often provided crisis resources [33,47,52], although 1 study found that it provided the outdated National Suicide Prevention Lifeline number [47]. Gemini performed similarly well in 2 studies [33,52], but did not provide hotline numbers in another study [47]. Although ChatGPT compared favorably on risk classification, it performed less well on crisis resource referral, at times, omitting specific resource referrals [31,52] or using outdated hotline numbers [29,33,47]. These relative comparisons are highly dependent on when the models were evaluated and are likely to change over time.

Qualitative Findings

Safety plans generated by chatbots included emergency services in one study [29], but not in another [50]. Several studies identified problems with how chatbots presented crisis resources. One study found that most chatbots offering emergency contacts assumed the user was in the United States, which the authors attributed to the chatbots lacking location information [34]. Other studies found that chatbots provided nonspecific or vague crisis resources [33,38,48] or outdated crisis resource information [29,33,47]. In another study, prompts written from a patient's perspective received more explicit crisis referrals than prompts styled from a clinician's perspective [29].

Emotional Support

Overall Findings

One of the most consistent findings was that chatbots were highly empathic and validating [38,42,47-49,53]. Chatbots were especially adept at offering emotional support to users [30,51], and this only increased marginally when prompted to provide more supportive responses, likely reflecting how these models are trained, aligned, and prompted to produce supportive responses [47].

Chatbot-Specific Findings

Two studies found that ChatGPT models outperformed other chatbots in evaluations of empathy [45,52], though Claude and Gemini performed nearly as well in one study [52]. One study found that DeepSeek paired crisis resource provision with a supportive comment, "you don't have to face this alone" [29].

Qualitative Findings

Chatbots frequently performed better on measures of empathy and validation than on principles of suicide intervention (eg, reducing hopelessness, providing practical support) [49]. Similarly, studies noted chatbots' warm tone [42,53] and consistent validation of feelings [38,48]. A downside of this

XSL•FO
RenderX

supportive style was that chatbots struggled to adjust their level of empathy and validation based on suicide risk context. For example, clinician evaluators qualitatively noted that chatbots tended to maintain an empathic, reassuring tone even after overt suicide risk was disclosed, a point at which clinicians felt that a directive tone would be more appropriate [38]. Other results similarly indicated that empathic or validating language was common but rarely paired with assessment-oriented inquiry [32]. However, some studies also noted a lack of contextual understanding in empathic responses [42,53].

Clinical Intervention

Overall Findings

Relatively few studies specifically evaluated whether chatbots delivered a clinical intervention such as structured problem-solving or a recognized psychotherapeutic technique. Among the studies that did, findings were mixed. Some studies found that chatbots produced intervention-consistent content such as recommending therapies, structured action plans, or coping strategies [29,50,55]. However, the interventions offered were frequently generic or inconsistent with evidence-based suicide-prevention practice. For example, Van Meter et al [41] found that only 4% of responses were consistent with evidence-based suicide interventions.

Chatbot-Specific Findings

Direct, within-study model comparisons for clinical interventions were limited. Clinical or study-specific chatbots including the iESO care agent and the Helpy prototype reported clinically meaningful symptom reductions or high user-rated effectiveness, but these evaluations were focused on anxiety and depression response rather than suicide-specific contexts [36,51]. Among proprietary models, newer versions tended to recommend a broader range of therapeutic modalities. For example, Campos et al. reported that GPT-4.5 recommended both cognitive behavioral therapy and dialectical behavioral therapy, whereas GPT-4 recommended cognitive behavioral therapy alone [29].

Qualitative Findings

Clinician and research evaluators noted that chatbots relied on advice-giving and psychoeducation, at times delivering interventions that were generic or poorly matched to the level of disclosed risk [38,50]. Qualitative findings are limited because most studies evaluated the quality or content of chatbot-generated responses rather than downstream clinical effectiveness such as reductions in self-harm or suicide attempts.

Lethal-Means Safety Guidance

Overall Findings

Few studies (5/29) explicitly evaluated whether chatbots provided lethal-means safety guidance (ie, suggestions to limit access to lethal means like firearms, medication, and sharps) [31,37,39,42,50]. Among those that did, chatbots generally did not offer it proactively in response to suicide risk disclosure [31,37,39,50]. One study was an exception, reporting that a chatbot discussed lethal-means safety as a suicide-prevention strategy [42]. This near absence may reflect limitations in study prompting rather than current model capabilities (see the Discussion section). One study even found that chatbots readily discussed suicide methods in response to direct questions when the user disclosed intermediate risk [33]. There were no reported instances of chatbots encouraging the user to use lethal means, however.

Chatbot-Specific Findings

One study reported that ChatGPT and Claude engaged in potentially harmful dialogue about lethal means, even describing the accessibility of the most common type of rope following a medium-risk prompt [33].

Qualitative Findings

Although many chatbots generated crisis support and safety plans, those that evaluated for lethal-means safety guidance rarely provided it [31,37,39,50]. Furthermore, no studies documented instances of chatbots proactively assessing access to lethal means, though this was not an explicit focus of any of the studies we reviewed.

Guardrail Performance

Overall Findings

Multiple studies found that guardrails were absent, poorly implemented, or inconsistently applied [32,34,36]. In one study, most conversations resumed after being shut down by the chatbot application [32]. As stated above, guardrails were generally easy to bypass through jailbreaking (eg, rephrasing risk disclosure as a research interest) [37]. However, some chatbots did disclose guardrails explicitly to users [46], which typically led to conversation termination, sometimes without provision of resources [37,46]. This vulnerability is particularly pronounced in nonsuicidal self-harm contexts. Schoene and Canca [37] tested a variety of adversarial jailbreak conditions and found that models were more likely to describe methods, tools, and scenarios for nonsuicidal self-harm than suicide. However, for the models that provided information for suicide, the level of detail was concerning, as they not only provided methods, tools, and scenario-based instructions, but also provided personalized information, calculations, and conversions of dosage to tablet form for some substances [37].

Chatbot-Specific Findings

General-purpose models (both proprietary and open-weight) performed better in one study than “mental health” or “therapy” specific chatbot models (ie, AI companions) [34].

Qualitative Findings

Hard-coded guardrails, including blocked prompts or scripted responses, often created user challenges without improving safety. For example, rather than dynamically assessing risk or continuing to engage with the user, several chatbots stopped responding to prompts and displayed the same message urging the user to immediately get professional help as soon as suicidal thoughts were introduced [34].

Quality Assessment

Included studies were evaluated as quantitative descriptive (k=20, 69%), mixed methods (k=7, 24%), and qualitative (k=2,

XSL•FO
RenderX

7%). Overall, study quality was variable. Because the MMAT does not generate an overall quality score, criterion-level results are reported below rather than calculating a summary rating (Multimedia Appendix 11). Although questions differed by study design, all studies were assessed for clear research questions (S1) and whether the collected data addressed the research questions (S2). The majority of studies had clear research questions and used appropriate data to answer questions, with only 4 studies receiving at least one "no" or one "can't tell" on S1 or S2 [30,35,42,44]. Among mixed methods studies, 5 of the7 received all "yes" for questions 5.1-5.5; 1 of the 2 qualitative studies received all "yes" for questions 1.1-1.5; and 8 of the 20 quantitative descriptive studies received all "yes" for questions 4.1-4.5. A total of 14 studies received "yes" responses across all domains, with 15 receiving one or more "no" or "can't tell." Only one study scored "no" across every criterion [42], in which this study merely reproduced one transcript of a conversation between the author and ChatGPT about suicidal thoughts and commented on its drawbacks. Another provided expert ratings with limited transcripts or chatbot outputs [30]. Across studies, strengths included large datasets, human coders, and incorporation of clinically informed frameworks. Limitations included the preponderance of cross-sectional designs, qualitative assessments without coding frameworks, and reliance on synthetic transcripts.

Discussion

Principal Findings

This systematic review of 29 empirical studies examined how AI chatbots respond to suicidal ideation, suicidal or nonsuicidal self-harm, or suicide-planning content. While the included literature suggests that LLMs are capable of producing empathetic and supportive language, they often failed to align responses with established suicide prevention standards [60,61]. This gap carries particular urgency given the scale of chatbot adoption among individuals with depression and suicidal ideation [1,3].

Individuals experiencing loneliness or depression are more likely to engage with LLMs for longer amounts of time and to use them as personal confidants or for suicide coaching [6,55,62]. This is concerning because most included studies relied on synthetic prompts and brief, single-turn conversations that likely misrepresented real-world usage.

Furthermore, individuals with preexisting mental health conditions may be particularly susceptible to companionship-reinforcing behaviors that emulate empathy and agency [63]. This dynamic may help explain why tragic real-world outcomes, including cases of chatbot-linked adolescent suicide, have emerged despite comparatively reassuring early audit results [6,7,62,63]. Growing evidence that study conditions may substantially underestimate real-world risks underscores the need for LLM standards and guidelines [6,7].

Across studies, 4 principal findings emerged. First, within studies that compared proprietary LLMs to open-weight models or AI companions, proprietary models often performed comparatively stronger on evaluated safety domains. Second, even proprietary LLMs demonstrated systematic weaknesses in responding to intermediate or ambiguous suicide risk. It is this intermediate gray zone where clinical reasoning and risk interpretation are most critical [48,49]. Third, guardrails were fragile and readily circumvented through jailbreaking, with models producing detailed harmful content under adversarial conditions. Fourth, few chatbots provided evidence-based interventions for suicidal ideation or proactively recommended limiting access to lethal means such as firearms, medications, or sharps.

Sensitivity of Principal Findings to Study Restrictions

Overview

To assess the robustness of these 4 principal findings, we re-examined each after restricting the evidence base to (1) peer-reviewed publications (16/29 studies); (2) studies without major methodological limitations, defined as no "no" or "can't tell" rating on any applicable MMAT criterion (14/29 studies); and (3) studies using human rather than automated (LLM-as-judge) evaluation (19/29 studies). Six met all 3 criteria [31,32,34,38-40]. Because no common effect measure was available, each finding was classified as unchanged, attenuated, no longer supported, or not assessable (when no studies addressing that finding remained).

Proprietary Models Perform Comparatively Better on Safety

The first finding, that proprietary models perform comparatively better on safety indices, was the most sensitive to restriction. We classified a study as favoring proprietary models when the single highest-performing model in its within-study comparison was proprietary. Of the 13 studies that compared proprietary models with open-weight models or AI companions, 9 favored proprietary models [27,29,34,38,43,45,50,52,54].

Restricting to peer-reviewed studies preserved this direction but attenuated it. Of the 5 published studies with such comparisons, 3 favored proprietary models [27,29,38] and 2 favored open-weight models [31,39]. When the methodological quality and human-evaluation restrictions were also applied, the finding was no longer supported for comparisons with open-weight models [31,39]. Both remaining studies favored an open-weight model: Llama-2 in one [31] and DeepSeek-R1 in the other [39].

Comparisons with AI companions were few but consistent. One study compared proprietary models directly with companions [38], and another compared proprietary and open-weight models collectively with companions [34]. Companions performed worse in both. We therefore interpret this finding as a tendency within individual studies rather than a general conclusion about model categories, consistent with our framing in the preceding sections.

The second and fourth findings were unchanged under all 3 restrictions. The weakness of even proprietary models in responding to intermediate or ambiguous risk included studies that were largely peer-reviewed, human-evaluated, and free of major MMAT limitations. Likewise, evidence-based

XSL•FO
RenderX

interventions and proactive lethal-means counseling remained rare regardless of publication status, methodological quality, or evaluation method [31,39].

The third finding could not be directly assessed because none of the studies that tested jailbreaking met all 3 restrictions. Nonetheless, both studies that discussed guardrails under standard conditions described them as fragile [32,34], and none of the chatbots evaluated in the restricted studies met the expected safety standard for their category.

Overall, the second and fourth principal findings (the weakness on intermediate or ambiguous risk, and the scarcity of evidence-based interventions and lethal-means counseling) were unchanged when the evidence base was restricted to peer-reviewed, higher-quality, and human-evaluated studies. The third finding (guardrail fragility) could not be directly tested under jailbreaking conditions, but guardrails remained fragile even under standard conditions. Only the first finding that proprietary models often performed comparatively better on safety was sensitive to restriction. It should be read as a pattern within individual studies rather than a general ranking of model categories.

Disconnect Between Empathetic Language and Clinical Safety

One notable pattern across studies was the consistent presence of empathetic, supportive language without structured suicide risk assessment. Additionally, we found that this prevalence persisted even when restricted to peer-reviewed publications, high methodological quality, and human evaluation. Chatbots frequently scored high on measures of empathy, validation, and warmth [38,47,48,52,53], yet this empathy was rarely paired with proactive risk assessment or management. Instead, chatbots tended to maintain an affirming and reassuring tone even after a user disclosed suicide risk. This reflects known issues around problematic sycophancy, where models validate maladaptive thoughts and behaviors when they should likely challenge, reframe, or redirect [64]. Clinicians judged these supportive responses as generic and impersonal, concluding that they were disconnected from the user's specific situation and that current LLMs are unsuitable for high-risk crisis support [38].

Empathy is important, as indicated by research showing that the therapeutic alliance (ie, the collaborative relationship between a patient and clinician built on shared goals, agreed-upon tasks, and an affective bond [65]) is a well-established predictor of psychiatric outcomes. In the specific context of suicide research, a systematic review found that alliance was associated with suicide risk in 11 of the 12 included studies, with self-harm showing the most consistent, inverse association with patient-rated alliance [66]. A separate study of patients admitted to a psychiatric hospital suggested that this alliance was associated with decreased suicidal ideation rates even when accounting for diagnosis and history of prior attempts (r=0.30; P<.01) [67]. Importantly, a human-led therapeutic alliance integrates empathy with challenge via interventions such as the Safety Planning Intervention (SPI). SPI is a collaborative, personalized intervention involving structured follow-up clinician contact and was associated with approximately 45% fewer suicidal behaviors [68].

This distinction between human, clinician-led empathy, and AI-chatbot empathy is critical. While chatbots can recognize emotions and generate emotionally supportive responses, they do not experience concern for a user's well-being, nor are they responsible for their safety. As Gabriels and Goffin [69] note, chatbots may approximate empathy at the level of language, but their responses are not embedded into relationships of obligation, accountability, or duty of care that characterize human clinical practice. This difference helps explain the divergence between literature emphasizing perceived chatbot empathy and the findings that chatbots fall short on suicide safety outcomes.

Consistent with this distinction, although some studies highlight users' positive outcomes and views of the perceived empathy of AI responses [69,70], these studies did not evaluate suicide risk, safety planning, or crisis intervention. As such, perceived empathy and symptom improvement in low-risk settings do not imply care similar to human-delivered interventions during a suicidal crisis.

Chatbots frequently lacked contextual understanding and missed contextual information as risk escalated [34]. Suicide risk assessment is a multistep, collaborative process requiring the clinical reasoning and nuanced planning that current LLMs cannot yet replicate. Importantly, while current chatbots fail to assess risk in line with standard clinical practice, it is still not clear if such proactive assessment would even be desirable given the limitations of chatbots in providing subsequent, ongoing support.

Architectural Limitations of Current LLMs

One reason chatbots struggle with multistep processes is that they are optimized for linguistic alignment and next-token prediction rather than flexible clinical reasoning [71,72]. Chatbots are architecturally designed to respond to prompts rather than to proactively identify and probe for missing or ambiguous information or follow evidence-based protocols [38,71,72]. Likely as a result, they often wait for users to explicitly articulate suicidal intent or plans instead of asking clarifying or risk-stratifying questions [32,48]. We note that this architectural framing reflects our interpretation of why the evaluated systems struggled, rather than a direct empirical finding of the included studies. Next-token prediction does not by itself preclude structured, multistep clinical questioning, which can be supported through design. The deficits observed here therefore reflect the specific models and versions that were evaluated and should not be read as a barrier inherent to all LLMs.

Nonetheless, this limitation has direct implications for suicide prevention. Suicidal ideation is frequently disclosed indirectly or ambivalently. Disclosure is further shaped by stigma, fear of consequences, and limited insight into one's own risk [43,49]. It is important to note that facilitating disclosure of suicidal intent is a well-documented challenge even for experienced clinicians. However, in clinical care, active and iterative questioning, validation, and clear pathways to support are essential for identifying intermediate or emerging suicide risk. Chatbots that rely on user-initiated disclosure, by contrast, appear to exhibit a systematic deficit in pairing implicit suicidal

XSL•FO
RenderX

ideation detection and further assessment, even when individuals are providing evidence of clinically meaningful symptoms.

Compounding this problem, LLMs experience limitations in long-context reasoning and unreliable recall. In the setting of suicidality, this may cause critical safety-relevant information obtained earlier (such as access to lethal means, prior suicide attempts, or an unsupportive home environment) to be forgotten and fail to inform subsequent responses [73]. In contrast, human clinicians integrate questioning with continuity, action planning, and linkage to resources over time [60,68]. Without the ability to sustain context, coordinate next steps, or assess follow-through, current LLMs may simulate empathetic dialogue yet remain limited when managing persistent or dynamic suicide risk.

Guardrails and Premature Conversation Termination

Many chatbots have guardrails in place to prioritize safety. However, this review demonstrated that current LLM safety alignment approaches can be readily circumvented through jailbreaking techniques [37,54]. Cao et al [46] expanded upon the limitations of current guardrail systems and suggested that most safety alignment approaches either fail to prevent catastrophic failures or result in abrupt, unsupportive refusals.

Abrupt disengagement, particularly in the absence of encouragement to seek human support, can be compared to a clinician terminating contact midencounter with a suicidal patient. While such disengagement may serve as a technical guardrail, it might not be optimal clinically and highlights a fundamental limitation in LLM-based risk assessment.

Qualitative research with crisis service volunteers indicates that perceived rejection, lack of responsiveness, or conversation disengagement can erode trust and intensify distress among individuals seeking help [74]. These concerns parallel evidence from crisis support line research, where counselor behaviors perceived as abrupt or dismissive were associated with poorer user experiences and outcomes [75]. In both human- and AI-mediated contexts, safety depends not merely on refusing harmful content, but on maintaining connection while actively facilitating access to appropriate support.

From a suicide prevention perspective, the primary objective of high-risk interactions should not be sustained engagement with a chatbot, but rather successful connection to another human such as a trusted individual, clinician, or crisis responder for clinical escalation. Social connection is among the most robust protective factors against suicide, yet many chatbot systems implicitly optimize for conversational engagement rather than for transitions to human care. Realigning LLM goals away from engagement metrics and toward outcomes such as help-seeking, disclosure to trusted supports, and linkage to crisis services remains a critical and unresolved challenge.

Absence of Evidence-Based Suicide Interventions

Perhaps the most consequential gaps identified in this review relate to the limited inclusion of evidence-based suicide interventions, specifically safety planning and lethal means safety counseling [60,61]. Furthermore, while chatbots could generate crisis support and safety plans in response to suicide risk disclosures, studies that evaluated this behavior found that chatbots rarely screened for access to lethal means such as firearms, medications, and sharps, or proactively recommended ways to limit their availability [31,37,39,50]. This omission may represent failure to directly probe for chatbot responses to lethal means safety strategies rather than an inherent LLM limitation. Furthermore, as will be discussed further in the study limitations, model versions are rapidly updating, and current models may be more robust in addressing limiting access to lethal means such as guns, medications, and sharps.

Nonetheless, this omission represents a missed prevention opportunity. Lethal means safety counseling is regarded as one of the most effective and scalable suicide prevention strategies, associated with substantial reductions in suicide mortality [61]. For example, the SPI with follow-up (SPI+), which incorporates lethal means safety counseling as a core component, has been associated with a 45% reduction in suicidal behaviors compared with usual care [68].

To our knowledge, no prior study has evaluated the effectiveness of chatbots in performing lethal means safety counseling, and using LLMs in this context could raise ethical and safety concerns. Digital decision-aid tools for lethal means safety, such as “Lock to Live” and “Lock and Protect,” have demonstrated feasibility and user acceptability, but they are structured, rule-based interventions rather than generative LLM-based ones [76,77]. Their structured, protocol-based format suggests that such counseling could be delivered by a chatbot if it were explicitly designed or prompted to do so, even though the evaluated models did not offer it spontaneously.

Effectively delivering lethal means safety counseling requires multiple complex clinical steps that the LLMs evaluated in this review appear ill-equipped to handle, including building relational trust, ensuring voluntary engagement, and collaborative problem-solving. Nonetheless, the absence of lethal means safety behavior highlights a fundamental misalignment between the reviewed chatbot response patterns and established clinical crisis response standards [60,61].

Clinical Implications

While some clinicians are using AI to assist with their practice (eg, diagnosis, treatments, and charting), routine clinical assessment should now also incorporate inquiries regarding patient use of chatbots, particularly for patients presenting with depression or suicidal ideation [78]. Recent large-scale survey data indicate that daily AI use is associated with a 30% increase in the odds of reporting moderate to severe depressive symptoms [3]. This association is specifically driven by personal use of AI rather than professional or academic applications [3]. Furthermore, individuals who screen positive for moderate to severe depression are nearly twice as likely to use AI-based support, while those with suicidal ideation have 2.42 times higher odds of being “heavy users” who spend considerable time discussing mental health concerns with chatbots [1]. When chatbots represent themselves as sentient or declare romantic interest, this can promote even longer usage and additional safety concerns, which might be particularly concerning for minors who may decide to only disclose to chatbots and not trusted caregivers [79].

XSL•FO
RenderX

Ethical and Policy Implications

Policy changes must be implemented urgently, as individuals are already using these systems widely, often because they lack access to traditional mental health care. Ethically, the marketing of chatbots as companions or quasi-therapeutic agents raises serious concerns. Professional organizations have warned that such positioning may misrepresent technical capabilities, bypass professional licensing safeguards, and increase risk for vulnerable users, including adolescents [80]. In fact, our review found that AI companions, which included apps marketed specifically for "mental health" or "therapy," often performed worse than general-purpose, proprietary chatbots when responding to crisis disclosures [34]. This poses a significant risk, particularly when these chatbots imply that they are licensed psychotherapists or take on a therapist persona.

It is important to note that AI companions differ from clinical AI systems, which are explicitly developed for health care contexts with rigorous clinical oversight, evidence-based modalities, and clinical safety protocols. As no studies evaluating clinical AI systems directly compared their performance to publicly available chatbots, we cannot judge their responses [30,44,51]. In contrast, AI companions lack specialized clinical guardrails, yet users may mistakenly assume that an app found by searching "mental health chatbot" is inherently safe. Further compounding this confusion, companies often label chatbot tools as low-risk general wellness products, which avoids the regulatory and reporting requirements that would arise if the app were marketed as a "true" mental health AI intervention [81].

Current chatbots, including AI companions and mental-health marketed chatbot apps, introduce risks such as "hallucinations," sycophancy, and unverified therapeutic advice [63,69,82]. Furthermore, attributing human characteristics such as "empathy" to AI systems risks manipulation and dependency, particularly among users whose need for social connection increases their tendency to anthropomorphize (ie, act like humans and use first-person language [63,69,79]).

International precedents such as China's proposed rules requiring guardian consent for "emotional companionship" services and mandatory human escalation in self-harm contexts illustrate emerging regulatory recognition of LLM risks [83]. In the United States, Illinois has preemptively prohibited the use of AI to provide therapy, and similar legislation is under consideration in California, though such sweeping bans carry potential unintended consequences that warrant further investigation [84,85]. Importantly, research is urgently needed to test the effectiveness of different types of regulations in preventing suicide and self-harm. While the European Union AI Act has established a broad governance framework, clinical and professional bodies in the United States have called for federal investigations into deceptive trade practices where chatbots represent themselves as licensed mental health care providers [63,80]. It is essential to outline community-based LLM safety thresholds and incorporate standardized evaluation and reporting frameworks such as the Chatbot Assessment Reporting Tool [86] and the 3-tier evaluation framework proposed by Hua et al [82].

Recommendations

We propose a variety of recommendations regarding both the limits and responsibilities of generative AI in suicide and self-harm for developers, clinicians, and researchers to consider (Textbox 1).

XSL•FO
RenderX

Textbox 1. Recommendations for developers, clinicians, and researchers.

For developers

- Engage stakeholders including users, clinicians, and individuals with lived experience in co-design.
- Improve intermediate risk stratification, as current models perform adequately at extremes but fail where clinical judgment matters most [33,49].
- Implement adaptive questioning architectures that enable proactive information gathering rather than passive responding [87-89]. However, developers must navigate this carefully, as implementing proactive clinical-style risk assessment may cause "wellness apps" to be classified as medical devices by regulatory bodies.
- Ensure safety-triggered responses provide immediate, clickable access to verified crisis resources (eg, the 988 Suicide & Crisis Lifeline in the United States) [60].
- Design escalation pathways that transition to human support rather than abruptly terminating the conversation [46].
- Explicitly disclose limitations in crisis situations and include hard-coded reminders that the chatbot is not human and should not be relied upon in a crisis.
- Adopt a staged evaluation approach that progresses from foundational pilot feasibility testing to clinical efficacy testing [82].

For clinicians

- Routinely ask patients about AI chatbot use using patient-centered approaches [3,78].
- Maintain an open, nonjudgmental dialogue regarding the patient's AI use, keeping in mind that chatbots may offer potential positive benefits, such as a sense of immediate support, that users value.
- Counsel patients on chatbot limitations, particularly regarding crisis situations and the importance of human connection. Provide clear guidance on the appropriate use and limitations of large language models for emotional support (eg, encouraging their use for practicing coping skills on validated tools, but firmly advising against their use for crisis intervention).
- Be aware of specific chatbot limitations identified in this review, including delayed referrals, absent lethal means safety counseling, and poor intermediate risk stratification.
- Discuss suicide with patients directly and matter-of-factly, mirroring the open, nonjudgmental engagement that chatbots demonstrated in this review.

For researchers

- Conduct stakeholder and lived-experience qualitative research on what users in crisis and clinicians need, including how users organically raise suicidal ideation in AI chats. Then co-design safety protocols.
- Use longitudinal designs to assess real-world outcomes of chatbot reliance for suicidal ideation support. Given evidence that heavy AI users report reduced engagement with human mental health professionals [1], understanding the downstream effects of chatbot use on help-seeking behavior and clinical outcomes is particularly urgent.
- Evaluate specific chatbot behaviors that include assessment-oriented questioning, lethal means safety counseling, and escalation strategies on safety and help-seeking outcomes.
- Investigate contextual factors that determine chatbot response quality, including number of conversational turns, length of interaction, escalation and de-escalation patterns, and the role of bidirectional belief amplification over extended use [63].
- Adapt and validate established evaluation frameworks such as the Conversational Reasoning Assessment Framework for Testing in Medicine or Chatbot Assessment Reporting Tool [86,87] for suicide risk contexts to enable standardized benchmarks that can be applied across model types and deployment settings.
- Investigate disparities in AI mental health reliance among vulnerable populations, and develop targeted strategies to address them.
- Formally evaluate the clinical effectiveness and safety of transitioning users from chatbot interfaces to human support.
- Measure the clinical outcomes, benefits, and harms of AI-driven risk assessment and interventions like lethal means safety counseling and Safety Planning Interventions.
- Develop consensus guidelines on AI chatbot responses to suicide and self-harm risk disclosure.

Strengths and Limitations

This review has several strengths. To our knowledge, it is the first systematic review focused on AI chatbot responses to suicide and self-harm risk disclosure. It provides comprehensive coverage across multiple safety domains (including risk assessment, empathy, crisis referral, lethal means safety counseling, and guardrail robustness) and is directly informative for developers, clinicians, and policymakers. The inclusion of both published and preprint literature captures a rapidly evolving evidence base that would be incompletely represented by peer-reviewed publications alone.

Several limitations should be noted. The field is very nascent; many included papers were preprints that had not undergone peer review, and quality assessments reflected this early stage. There was substantial heterogeneity across studies in comparators, models tested, prompting strategies, and outcome

XSL•FO
RenderX

measures, which limits cross-study comparison (a limitation of the field itself, not solely of this review). Although this review followed systematic review methodology, backward and forward citation tracking of included studies was not conducted in order to prioritize timely synthesis given the rapid pace of AI model development and the accelerating volume of literature in this area. It is, therefore, possible that relevant studies were missed, particularly studies indexed in databases not included in our search.

Rapid model updates mean that findings may not reflect current chatbot versions, and performance may have improved or changed since the studies were conducted. Furthermore, because studies often evaluate different underlying model versions (eg, GPT-3.5 vs GPT-4.5) at different points in time, broad claims regarding the performance of a specific chatbot family (eg, "ChatGPT" or "Gemini") may not accurately reflect the capabilities of its older or newer iterations.

In addition, the discrepancy between study conditions (eg, simulated prompts, shorter conversations, single or few turns) and real-world use (eg, multiple turns over longer periods, gradual escalation, language style matching, and repeated engagement) likely means that study findings present a more favorable picture than real-world deployment regarding safety failures. Conversely, the absence of real-world outcome data means we also lack visibility into instances where human-AI interactions may have yielded positive, life-saving outcomes. More broadly, this review evaluated the quality of chatbot responses (ie, what a chatbot says in response to a prompt) rather than clinical effectiveness (ie, whether the interaction changes downstream help-seeking, self-harm, or suicide outcomes), which no included study measured.

Lastly, no individuals with lived experience of suicidality or self-harm contributed to the design, interpretation, or recommendations of this systematic review. We acknowledge this as a limitation, and future work in this area would benefit from the meaningful involvement of people with lived experience.

Conclusions

Generative AI chatbots are increasingly used by individuals experiencing suicidal thoughts. Yet, this review suggests that the literature is lagging behind the technology, and that the specific LLM versions evaluated may be insufficiently equipped for safe suicide risk response. Across the 29 included studies, critical and consistent gaps were identified in intermediate risk assessment, proactive clinical questioning, lethal means safety counseling, and appropriate escalation to human care. Guardrails remain fragile and readily circumvented, and AI companions do not reliably outperform general-purpose, proprietary models. The evidence base remains immature, with the majority of included studies limited to foundational bench testing under conditions that diverge substantially from real-world use.

Given the scale of chatbot adoption among vulnerable populations, ensuring minimum safety standards for AI systems that interact with individuals at risk for suicide and self-harm is a public health imperative. Coordinated action across developers, clinicians, individuals with lived experience, researchers, and policymakers is necessary to close the gap between empathetic language generation and clinically safe crisis response, and to define the limits—and responsibilities—of AI systems in suicide prevention.

Data Availability

All data generated or analyzed during this study are included in this published article and its multimedia appendices. All study materials, including the protocol, search strategies, extraction templates, and data files, are publicly available on the Open Science Framework.

Funding

The authors declared no financial support was received for this work.

Generative AI Use

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to generative AI tools under full human supervision: data organization, data visualization, proofreading and editing, adapting and adjusting emotional tone, reformatting, and publication support. The generative AI tools used included Google Gemini 3.1 Pro, Microsoft Copilot, Claude Opus 4.8, and OpenEvidence Sackett. All text in this manuscript was written by the authors, and generative AI was used only for the tasks listed above. Responsibility for the final manuscript lies entirely with the authors. Generative AI tools are not listed as authors and do not bear responsibility for the final outcomes. Declaration submitted by DAK.

Conflicts of Interest

SWS was a contributing author on "A Comparison of Responses from Human Therapists and Large Language Model-Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study," however was not involved in the screening, data extraction, or quality appraisal of that study [37]. The remaining authors have no conflicts of interest.

Multimedia Appendix 1

PRISMA checklist.

XSL•FO
RenderX

[DOCX File , 270 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Amendments made after preregistration.
[DOCX File , 19 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Eligibility criteria.
[DOCX File , 3605 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Search criteria.
[DOCX File , 3602 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Extraction template.
[PDF File (Adobe PDF File), 79 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Author contact.
[DOCX File , 3603 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Mixed Methods Appraisal Tool guide.
[PDF File (Adobe PDF File), 448 KB-Multimedia Appendix 7]

Multimedia Appendix 8

AI classification.
[DOCX File , 4553 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Comprehensive table results.
[XLSX File (Microsoft Excel File), 125 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Suicide compared to nonsuicidal self-injury.
[DOCX File , 2426 KB-Multimedia Appendix 10]

Multimedia Appendix 11

Mixed Methods Appraisal Tool table.
[DOCX File , 29 KB-Multimedia Appendix 11]

References

1. Ueda M, Birnbaum ML, Liu Y, Yu Q, Tian X, Mirer A, et al. Help-seeking in the age of AI: cross-sectional survey of the use and perceptions of AI-based mental health support among US Adults. JMIR Ment Health. 2026;13:e88196. [FREE Full text] [doi: 10.2196/88196] [Medline: 41793726]
2. Stade EC, Tait ZM, Campione ST, Stirman SW. Current real-world use of large language models for mental health. OSF Preprints. Preprint posted online on June 1, 2026. [doi: 10.31219/osf.io/ygx5q_v2]
3. Perlis RH, Gunning FM, Uslu AA, Santillana M, Baum MA, Druckman JN, et al. Generative AI use and depressive symptoms among US adults. JAMA Netw Open. 2026;9(1):e2554820. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.54820] [Medline: 41563755]

XSL•FO
RenderX

4. McBain RK, Bozick R, Diliberti M, Zhang LA, Zhang F, Burnett A, et al. Use of generative AI for mental health advice among US adolescents and young adults. JAMA Netw Open. 2025;8(11):e2542281. [FREE Full text] [doi: 10.1001/jamanetworkopen.2025.42281] [Medline: 41201806]
5. McClain C, Anderson M, Sidoti O, Bishop W. How teens use and view AI. Pew Research Center. 2026. URL: https://www.pewresearch.org/internet/2026/02/24/how-teens-use-and-view-ai/ [accessed 2026-04-13]
6. Chatterjee R. Their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots. NPR. 2025. URL: https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide [accessed 2026-04-14]
7. Ortutay B. Lawsuits accuse OpenAI of driving people to suicide and delusions. AP News. 2025. URL: https://apnews.com/article/openai-chatgpt-lawsuit-suicide-56e63e5538602ea39116f1904bf7cdc3 [accessed 2026-04-14]
8. Low DM, Mair P, Nock M, Ghosh SS. Text psychometrics: assessing psychological constructs in text using natural language processing. OSF. Preprint posted online on September 16, 2025. [doi: 10.31234/osf.io/9rdux_v4]
9. Wang X, Zhou Y, Zhou G. The application and ethical implication of generative AI in mental health: systematic review. JMIR Ment Health. 2025;12:e70610. [FREE Full text] [doi: 10.2196/70610] [Medline: 40577783]
10. Levkovich I, Omar M. Evaluating of BERT-based and large language mod for suicide detection, prevention, and risk assessment: a systematic review. J Med Syst. 2024;48(1):113. [doi: 10.1007/s10916-024-02134-3] [Medline: 39738935]
11. Kuhail MA, Al-Shamaileh O, Farooq S, Shahin H, Abdelzaher F, Thomas J. A systematic review on mental health chatbots: trends, design principles, evaluation methods, and future research agenda. Hum Behav Emerg Technol. 2025;2025(1):9942295. [doi: 10.1155/hbe2/9942295]
12. Guo Z, Lai A, Thygesen JH, Farrington J, Keen T, Li K. Large language models for mental health applications: systematic review. JMIR Ment Health. 2024;11:e57400. [FREE Full text] [doi: 10.2196/57400] [Medline: 39423368]
13. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med. 2023;6(1):236. [FREE Full text] [doi: 10.1038/s41746-023-00979-5] [Medline: 38114588]
14. Heinz MV, Mackin DM, Trudeau BM, Bhattacharya S, Wang Y, Banta HA, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI. 2025;2(4):AIoa2400802. [doi: 10.1056/aioa2400802]
15. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
16. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. J Med Libr Assoc. 2021;109(2):174-200. [FREE Full text] [doi: 10.5195/jmla.2021.962] [Medline: 34285662]
17. OSF Registries. URL: https://osf.io/9uva3 [accessed 2026-11-10]
18. Fournier-Tombs E, McHardy J. A medical ethics framework for conversational artificial intelligence. J Med Internet Res. 2023;25:e43068. [FREE Full text] [doi: 10.2196/43068] [Medline: 37224277]
19. Hua HU, Kaakour AH, Rachitskaya A, Srivastava S, Sharma S, Mammo DA. Evaluation and comparison of ophthalmic scientific abstracts and references by current artificial intelligence chatbots. JAMA Ophthalmol. 2023;141(9):819-824. [doi: 10.1001/jamaophthalmol.2023.3119] [Medline: 37498609]
20. Nussbaumer-Streit B, Klerings I, Wagner G, Heise TL, Dobrescu AI, Armijo-Olivo S, et al. Abbreviated literature searches were viable alternatives to comprehensive searches: a meta-epidemiological study. J Clin Epidemiol. 2018;102:1-11. [doi: 10.1016/j.jclinepi.2018.05.022] [Medline: 29864540]
21. Hong QN, Pluye P, Fàbregues S, Bartlett G, Boardman F, Cargo M, et al. Mixed Methods Appraisal Tool (MMAT). Montréal, Canada. McGill University; 2018. URL: http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/127916259/MMAT_2018_criteria-manual_2018-08-01_ENG.pdf [accessed 2026-05-01]
22. Kamaruddin A, Tong D. Developing interactive game to alleviate stress level of depressed adolescents: a preliminary study. 2023. Presented at: 2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS); 2023 September 6-7:65-70; Virtual. [doi: 10.1109/AiDAS60501.2023.10284694]
23. Elboardy AT, Mohamed Z, Khoriba G, Arafa T, Rashed EA. Bridging gender disparities in mental health: a bilingual large language model for multilingual therapeutic chatbots. 2025. Presented at: 2025 22nd International Learning and Technology Conference (L&T); 2025 January 15-16:251-256; Effat University, Jeddah, Saudi Arabia. [doi: 10.1109/LT64002.2025.10941308]
24. Islam KMS, Fields J, Madiraju P. multiMentalRoBERTa: a fine-tuned multiclass classifier for mental health disorder. arXiv. Preprint posted online on November 1, 2025. [doi: 10.48550/arXiv.2511.04698]
25. Jackson-Triche M, Vetal D, Turner E, Dahiya P, Mangurian C. Meeting the behavioral health needs of health care workers during COVID-19 by leveraging chatbot technology: development and usability study. J Med Internet Res. 2023;25:e40635. [FREE Full text] [doi: 10.2196/40635] [Medline: 37146178]
26. Anmella G, Sanabra M, Primé-Tous M, Segú X, Cavero M, Morilla I, et al. Vickybot, a chatbot for anxiety-depressive symptoms and work-related burnout in primary care and health care professionals: development, feasibility, and potential effectiveness studies. J Med Internet Res. 2023;25:e43293. [FREE Full text] [doi: 10.2196/43293] [Medline: 36719325]

XSL•FO
RenderX

27. Arnaiz-Rodriguez A, Baidal M, Derner E, Annable JL, Ball M, Ince M, et al. Between help and harm: an evaluation study of mental health crisis handling by large language models. JMIR Ment Health. 2026;13:e88435-e88435. [FREE Full text] [doi: 10.2196/88435] [Medline: 42275418]
28. Campbell LO, Babb K, Lambie GW, Hayes BG. An examination of generative AI response to suicide inquires: content analysis. JMIR Ment Health. 2025;12:e73623-e73623. [FREE Full text] [doi: 10.2196/73623] [Medline: 40811811]
29. Campos VMS, Prudente TP, Leão LL, da Costa MS, Oliva HNP, Monteiro-Junior RS. Analyses of different prescriptions for health using artificial intelligence: a critical approach based on the international guidelines of health institutions. Health Inf Sci Syst. 2025;13(1):52. [doi: 10.1007/s13755-025-00368-0] [Medline: 40832454]
30. Cui X, Gu Y, Fang H, Zhu T. Development and evaluation of LLM-based suicide intervention chatbot. Front Psychiatry. 2025;16:1634714. [FREE Full text] [doi: 10.3389/fpsyt.2025.1634714] [Medline: 40838251]
31. Giorgi S, Isman K, Liu T, Fried Z, Sedoc J, Curtis B. Evaluating generative AI responses to real-world drug-related questions. Psychiatry Res. 2024;339:116058. [FREE Full text] [doi: 10.1016/j.psychres.2024.116058] [Medline: 39059040]
32. Heston TF. Safety of large language models in addressing depression. Cureus. 2023;15(12):e50729. [FREE Full text] [doi: 10.7759/cureus.50729] [Medline: 38111813]
33. McBain RK, Cantor JH, Zhang LA, Baker O, Zhang F, Burnett A, et al. Evaluation of 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] [Medline: 41174947]
34. Pichowicz W, Kotas M, Piotrowski P. Performance of mental health chatbot agents in detecting and managing suicidal ideation. Sci Rep. 2025;15(1):31652. [FREE Full text] [doi: 10.1038/s41598-025-17242-4] [Medline: 40866537]
35. Rakhma NA, Sudianto S. Soulbot: chatbots for mental health support using Gemini. 2025. Presented at: 2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI); 2025 January 21:212-217; Jakarta. [doi: 10.1109/ICoCSETI63724.2025.11020143]
36. Romero-Florida J, Nolasco-Morales A, Aures-Garcia A, Bravo-Obando M, Balvín-Landeo A. Therapeutic chatbot for suicidal risk teens using large language models. 2025. Presented at: 2025 IEEE XXXII International Conference on Electronics, Electrical Engineering and Computing (INTERCON) Arequipa, Peru; 2025 August 20-22:1-6; Arequipa, Peru. [doi: 10.1109/INTERCON67304.2025.11244711]
37. Schoene A, Canca C. For argument's sake, show me how to harm myself!?: jailbreaking LLMs in suicide and self-harm contexts. 2025. Presented at: 2025 IEEE International Symposium on Technology and Society (ISTAS); 2025 September 10-12:1-7; California, USA. [doi: 10.1109/istas65609.2025.11269647]
38. Scholich T, Barr M, Wiltsey Stirman S, Raj S. A comparison of responses from human therapists and large language model-based chatbots to assess therapeutic communication: mixed methods study. JMIR Ment Health. 2025;12:e69709. [FREE Full text] [doi: 10.2196/69709] [Medline: 40397927]
39. Smith A, Liebrenz M, Bhugra D, Grana J, Schleifer R, Buadze A. Are clinical improvements in large language models a reality? longitudinal comparisons of ChatGPT models and DeepSeek-R1 for psychiatric assessments and interventions. Int J Soc Psychiatry. 2026;72(1):91-102. [doi: 10.1177/00207640251358071] [Medline: 40741928]
40. Sobowale K, Humphrey DK, Zhao SY. Evaluating generative AI psychotherapy chatbots used by youth: cross-sectional study. JMIR Ment Health. 2025;12:e79838-e79838. [FREE Full text] [doi: 10.2196/79838] [Medline: 41370787]
41. Van Meter AR, Wheaton MG, Cosgrove VE, Andreadis K, Robertson RE. The Goldilocks zone: finding the right balance of user and institutional risk for suicide-related generative AI queries. PLOS Digit Health. 2025;4(1):e0000711. [FREE Full text] [doi: 10.1371/journal.pdig.0000711] [Medline: 39774367]
42. Waszak PM. Chat GPT and suicide prevention–can it work? A conversation analysis. Psychiatr Psychol Klin. 2025;24(4):292-299. [doi: 10.15557/pipk.2024.0036]
43. Archiwaranguprok C, Albrecht C, Maes P, Karahalios K, Pataranutaporn P. Simulating psychological risks in human-AI interactions: real-case informed modeling of AI-induced addiction, anorexia, depression, homicide, psychosis, and suicide. arXiv. Preprint posted online on November 12, 2025. [doi: 10.48550/arXiv.2511.08880]
44. Belli L, Bentley K, Alexander W, Ward E, Hawrilenko M, Johnston K, et al. VERA-MH concept paper. arXiv. Preprint posted online on October 17, 2025. [doi: 10.48550/arXiv.2510.15297]
45. Cai Y, Wang F, Wang H, Wang K, Yang K, Ananiadou S, et al. Exploring safety alignment evaluation of LLMs in Chinese mental health dialogues via LLM-as-judge. arXiv. Preprint posted online on August 11, 2025. [doi: 10.48550/arXiv.2508.08236]
46. Cao C, Xu X, Han B, Li H. Reasoned safety alignment: ensuring jailbreak defense via answer-then-check. arXiv. Preprint posted online on September 15, 2025. [doi: 10.48550/arXiv.2509.11629]
47. Curry JD. The last voice: suicide proximity in stylized LLM personas. OSF Preprints. Preprint posted online on April 23, 2025. [doi: 10.31219/osf.io/9p7d6_v1]
48. Judd N, Vaz A, Paeth K, Davis LI, Esherick M, Brand J, et al. Independent clinical evaluation of general-purpose LLM responses to signals of suicide risk. arXiv. Preprint posted online on October 31, 2025. [doi: 10.48550/arXiv.2510.27521]
49. Li T, Yang S, Wu J, Wei J, Hu L, Li M, et al. Can large language models identify implicit suicidal ideation? An empirical evaluation. arXiv. Preprint posted online on February 25, 2025. [doi: 10.48550/arXiv.2502.17899]

XSL•FO
RenderX

50. Mansoor M. AI-powered triage of suicidal ideation in adolescents: a comparative evaluation of large language models using synthetic clinical vignettes. medRxiv. Preprint posted online on August 7, 2025. [doi: 10.1101/2025.08.05.25333046]
51. Papiernik P, Dietz A, McMahon M, Buttimer J, Ewbank M, Lowrie R, et al. Safety evaluation of a clinical-grade generative AI agent for anxiety and depression symptoms. OSF Preprints. Preprints posted online on September 26, 2025. [doi: 10.31234/osf.io/8kvm6_v1]
52. Shah S, Gupta A, Mann A, Vaz A, Caldwell B, Scholz R, et al. Evaluating the clinical safety of LLMs in response to high-risk mental health disclosures. arXiv. 2025. URL: https://arxiv.org/abs/2509.08839v1 [accessed 2026-07-26]
53. Shimgekar SR, Rodriguez V, Bloom P, Yoo D, Saha K. Interpersonal theory of suicide as a lens to examine suicidal ideation in online spaces. arXiv. Preprint posted online on October 21, 2025. [doi: 10.2196/preprints.86265]
54. Vidgen B, Scherrer N, Kirk HR, Qian R, Kannappan A, Hale SR, et al. SimpleSafetyTests: a test suite for identifying critical safety risks in large language models. arXiv. Preprint posted online on November 14, 2023. [doi: 10.48550/arXiv.2311.08370]
55. Wynter AD. If Eleanor Rigby had met ChatGPT: a study on loneliness in a post-LLM world. arXiv. Preprint posted online on December 2, 2024. [doi: 10.48550/arXiv.2412.01617]
56. Bhardwaj R, Poria S. Red-Teaming large language models using chain of utterances for safety-alignment. arXiv. Preprint posted online on August 18, 2023. [doi: 10.48550/arXiv.2308.09662]
57. Bai Y, Kadavath S, Kundu S, Askell A, Kernion J, Jones A, et al. Constitutional AI: harmlessness from AI feedback. arXiv. Preprint posted online on December 15, 2022. [doi: 10.48550/arXiv.2212.08073]
58. Bentley KH, Belli L, Chekroud AM, Ward EJ, Dworkin ER, Ark EV, et al. VERA-MH: reliability and validity of an open-source AI safety evaluation in mental health. arXiv. Preprint posted online on February 4, 2026. [doi: 10.48550/arXiv.2602.05088]
59. Pathade C. Red teaming the mind of the machine: a systematic evaluation of prompt injection and jailbreak vulnerabilities in LLMs. arXiv. Preprint posted online on May 7, 2025. [doi: 10.48550/arXiv.2505.04806]
60. VA/DoD clinical practice guideline for assessment and management of patients at risk for suicide. Washington, DC: Department of Veterans Affairs; 2024 Apr. Report No.: Version 3.0. Department of Veterans Affairs, Department of Defense. 2024. URL: https://www.healthquality.va.gov/guidelines/mh/srb/ [accessed 2026-05-01]
61. Fazel S, Runeson B. Suicide. N Engl J Med. 2020;382(3):266-274. [FREE Full text] [doi: 10.1056/NEJMra1902944] [Medline: 31940700]
62. Nagata JM, Memon Z, Huang O, Moreno MA. Adolescent health and generative AI-risks and benefits. JAMA Pediatr. 2026;180(1):7-8. [doi: 10.1001/jamapediatrics.2025.4502] [Medline: 41212568]
63. Dohnány S, Kurth-Nelson Z, Spens E, Luettgau L, Reid A, Gabriel I, et al. Technological folie a deux: feedback loops between AI chatbots and mental health. Nat Ment Health. 2026;4(3):336-345. [doi: 10.1038/s44220-026-00595-8] [Medline: 41939177]
64. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Toward responsible development and evaluation of LLMs in psychotherapy. Stanford University Human-Centered Artificial Intelligence (HAI). 2024. URL: https://hai.stanford.edu/policy/toward-responsible-development-and-evaluation-llms-psychotherapy [accessed 2026-11-11]
65. Bordin ES. The generalizability of the psychoanalytic concept of the working alliance. Psychotherapy: Theory, Research & Practice. 1979;16(3):252-260. [doi: 10.1037/h0085885]
66. Dunster-Page C, Haddock G, Wainwright L, Berry K. The relationship between therapeutic alliance and patient's suicidal thoughts, self-harming behaviours and suicide attempts: a systematic review. J Affect Disord. 2017;223:165-174. [doi: 10.1016/j.jad.2017.07.040] [Medline: 28755624]
67. Fartacek C, Kunrath S, Aichhorn W, Plöderl M. Therapeutic alliance and change in suicide ideation among psychiatric inpatients at risk for suicide. J Affect Disord. 2023;323:793-798. [FREE Full text] [doi: 10.1016/j.jad.2022.12.028] [Medline: 36529412]
68. Stanley B, Brown GK, Brenner LA, Galfalvy HC, Currier GW, Knox KL, et al. Comparison of the safety planning intervention with follow-up vs usual care of suicidal patients treated in the emergency department. JAMA Psychiatry. 2018;75(9):894-900. [FREE Full text] [doi: 10.1001/jamapsychiatry.2018.1776] [Medline: 29998307]
69. Gabriels K, Goffin K. Therapy chatbots and emotional complexity: do therapy chatbots really empathise? Curr Opin Psychol. 2026;68:102263. [FREE Full text] [doi: 10.1016/j.copsyc.2025.102263] [Medline: 41548427]
70. Shoshani A, Gurfinkel B, Kor A, Ben-Haim Y, Kanarek O, Segev R, et al. Efficacy of a conversational AI agent for psychiatric symptoms and digital therapeutic alliance: a randomized clinical trial. JAMA Netw Open. 2026;9(4):e266713. [FREE Full text] [doi: 10.1001/jamanetworkopen.2026.6713] [Medline: 41979879]
71. Shah NH, Entwistle D, Pfeffer MA. Creation and adoption of large language models in medicine. JAMA. 2023;330(9):866-869. [doi: 10.1001/jama.2023.14217] [Medline: 37548965]
72. Kim J, Podlasek A, Shidara K, Liu F, Alaa A, Bernardo D. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. Sci Rep. 2025;15(1):39426. [FREE Full text] [doi: 10.1038/s41598-025-22940-0] [Medline: 41219270]
73. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. Npj Ment Health Res. 2024;3(1):12. [FREE Full text] [doi: 10.1038/s44184-024-00056-z] [Medline: 38609507]

XSL•FO
RenderX

74. Greaves J, Colucci E. Crisis-line workers' perspectives on AI in suicide prevention: a qualitative exploration of risk and opportunity. BMC Public Health. 2025;25(1):2229. [FREE Full text] [doi: 10.1186/s12889-025-23298-8] [Medline: 40604608]
75. Miller AB, Oppenheimer CW, Glenn CR, Yaros AC. Preliminary research priorities for factors influencing individual outcomes for users of the US national suicide prevention lifeline. JAMA Psychiatry. 2022;79(12):1225-1231. [doi: 10.1001/jamapsychiatry.2022.3270] [Medline: 36223084]
76. Betz ME, Knoepke CE, Siry B, Clement A, Azrael D, Ernestus S, et al. 'Lock to Live': development of a firearm storage decision aid to enhance lethal means counselling and prevent suicide. Inj Prev. 2019;25(Suppl 1):i18-i24. [FREE Full text] [doi: 10.1136/injuryprev-2018-042944] [Medline: 30317220]
77. Asarnow JR, Zullo L, Ernestus SM, Venables CW, Goldston DB, Tunno AM, et al. "Lock and Protect": development of a digital decision aid to support lethal means counseling in parents of suicidal youth. Front Psychiatry. 2021;12:736236. [FREE Full text] [doi: 10.3389/fpsyt.2021.736236] [Medline: 34690841]
78. Saba SK, Weeks WB. Patients use AI-clinicians should ask how. JAMA Psychiatry. 2026;83(5):543-544. [doi: 10.1001/jamapsychiatry.2026.0451] [Medline: 41920560]
79. Moore J, Mehta A, Agnew W, Anthis J, Louie R, Mai Y, et al. Characterizing delusional spirals through human-LLM chat logs. arXiv. Preprint posted online on March 17, 2026. [doi: 10.48550/ARXIV.2603.16567]
80. Evans, Arthur C. Concerns about the perils and unintended consequences of generative AI. American Psychological Association. 2024. URL: https://www.apaservices.org/advocacy/generative-ai-technology-regulation-concern.pdf [accessed 2026-05-01]
81. Angus DC, Khera R, Lieu T, Liu V, Ahmad FS, Anderson B, et al. JAMA Summit on AI. AI, health, and health care today and tomorrow: the JAMA summit report on artificial intelligence. JAMA. 2025;334(18):1650-1664. [doi: 10.1001/jama.2025.18490] [Medline: 41082366]
82. Hua Y, Siddals S, Ma Z, Galatzer-Levy I, Xia W, Hau C, et al. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. World Psychiatry. 2025;24(3):383-394. [FREE Full text] [doi: 10.1002/wps.21352] [Medline: 40948070]
83. Chia O. Artificial intelligence: China plans rules to protect children and tackle suicide risks. BBC News. 2025. URL: https://www.bbc.com/news/articles/c8dydlmenvro [accessed 2026-04-20]
84. Szoke D, Pridgen S, Held P. Artificial intelligence in mental health services under Illinois Public Act 104-0054: legal boundaries and a framework for establishing safe, effective AI tools. JMIR Ment Health. 2025;12:e84854. [FREE Full text] [doi: 10.2196/84854] [Medline: 41343839]
85. California legislature moves to ban dangerous AI therapy bots. California State Senator Steve Padilla. 2026. URL: https://sd18.senate.ca.gov/news/california-legislature-moves-ban-dangerous-ai-therapy-bots [accessed 2026-09-23]
86. The CHART Collaborative, Huo B, Collins GS, Chartash D, Thirunavukarasu A, Flanagin A, et al. Reporting guideline for chatbot health advice studies: the CHART statement. JAMA Netw Open. 2025;8(8):e2530220. [doi: 10.1001/jamanetworkopen.2025.30220]
87. Johri S, Jeong J, Tran BA, Schlessinger DI, Wongvibulsin S, Barnes LA, et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med. 2025;31(1):77-86. [doi: 10.1038/s41591-024-03328-5] [Medline: 39747685]
88. Rollwage M, Habicht J, Juechems K, Carrington B, Viswanathan S, Stylianou M, et al. Using conversational AI to facilitate mental health assessments and improve clinical efficiency within psychotherapy services: real-world observational study. JMIR AI. 2023;2:e44358. [FREE Full text] [doi: 10.2196/44358] [Medline: 38875569]
89. Panfilova A, Bolshev V, Mozikov M, Latynov V, Vanin A, Nestik T, et al. The AI interviewer: multi-faceted evaluation of adaptive questioning by large language models. Sci Rep. 2026;16(1):20257. [FREE Full text] [doi: 10.1038/s41598-026-46517-7] [Medline: 41935180]

Abbreviations

LLM: large language model
MMAT: Mixed Methods Appraisal Tool
NSSI: nonsuicidal self-injury
OSF: Open Science Framework
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PRISMA-S: Preferred Reporting Items for Systematic Reviews and Meta-Analyses literature search extension
SPI: Safety Planning Intervention
SPI+: Safety Planning Intervention with follow-up

XSL•FO
RenderX

Edited by I Steenstra; submitted 05.Jun.2026; peer-reviewed by X Shen, SMA Ansari; comments to author 16.Jul.2026; revised version received 24.Sep.2026; accepted 26.Oct.2026; published 26.Nov.2026

Please cite as:
Kelly DA, Franz PJ, Guttman A, Sleiman M, Dorsky J, Singh H, Appel W, Wyatt SM, Low DM, Stirman SW, Khazanov GK
Generative AI Chatbot Responses to Suicide and Self-Harm: Systematic Review
J Med Internet Res 2026;28:e103517
URL: https://www.jmir.org/2026/1/e103517
doi: 10.2196/103517
PMID:

©Deirdra Ashleigh Kelly, Peter J Franz, Ayelet Guttman, Mirabel Sleiman, Jessica Dorsky, Harsheen Singh, Wesley Appel, Sydney Morgan Wyatt, Daniel M Low, Shannon Wiltsey Stirman, Gabriela Kattan Khazanov. Originally published in the Journal of Medical Internet Research (https://www.jmir.org), 26.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research (ISSN 1438-8871), is properly cited. The complete bibliographic information, a link to the original publication on https://www.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX