International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	SUICIDE AND CRISIS RISK DETECTION IN MENTAL HEALTH CHATBOTS: TECHNICAL APPROACHES, SAFETY EVIDENCE, AND GOVERNANCE CHALLENGES
DOI	https://doi.org/10.31435/ijitss.3(51).2026.6484
RECEIVED	24 June 2026
ACCEPTED	14 September 2026
PUBLISHED	16 September 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

SUICIDE AND CRISIS RISK DETECTION IN MENTAL HEALTH CHATBOTS: TECHNICAL APPROACHES, SAFETY EVIDENCE, AND GOVERNANCE CHALLENGES

Kacper Wrzosek *(Corresponding Author, Email: kacperwrzosek2003@gmail.com)*
Student Research Circle at the Chair and Department of Epidemiology and Clinical Research Methodology, Medical University of Lublin, 20-093 Lublin, Poland
ORCID ID: 0009-0009-6215-2134

Karina Udrycka
Medical University of Lodz, 90-419 Lodz, Poland
ORCID ID: 0009-0003-8577-4100

Kornelia Kuchta
Student Research Circle at the Chair and Department of Epidemiology and Clinical Research Methodology, Medical University of Lublin, 20-093 Lublin, Poland
ORCID ID: 0009-0008-0272-5968

Jan Poleszak
Student Research Circle at the Chair and Department of Epidemiology and Clinical Research Methodology, Medical University of Lublin, 20-093 Lublin, Poland
ORCID ID: 0009-0001-4265-0998

Maria Leśniak
Student Research Circle at the Chair and Department of Epidemiology and Clinical Research Methodology, Medical University of Lublin, 20-093 Lublin, Poland
ORCID ID: 0009-0007-2413-6112

Piotr Kowalewski
Student Research Circle at the Chair and Department of Epidemiology and Clinical Research Methodology, Medical University of Lublin, 20-093 Lublin, Poland
ORCID ID: 0009-0002-6687-2226

ABSTRACT

Mental health chatbots built on large language models are now used at population scale, including by adolescents disclosing suicidal thoughts, yet the evidence base supporting them was generated to measure symptom change rather than crisis safety. This narrative review synthesises 43 records published between 2021 and 2026 and identified through PubMed, Scopus and Web of Science, organised around technical approaches to suicide and crisis risk detection, empirical safety evidence, and governance responses. Detection from natural language is technically feasible: classifiers applied to crisis conversation corpora reach areas under the curve between 0.89 and 0.92, and prompt-defined thresholds can reduce false negatives to zero at sub-second latency, although only at the cost of substantially elevated false alarms. Reported performance is not comparable across studies because label provenance varies, and model errors concentrate on the cases about which expert clinicians themselves disagree. Deployed systems perform poorly: among 29 commercial applications tested against escalating suicidal risk scenarios, none met the criteria for an adequate response, and general-purpose models align with clinical judgement at the extremes of the risk continuum but not in the intermediate range where assessment is most consequential. Where crisis handling has demonstrably worked, detection was separated from the conversational model and coupled to a human escalation pathway. Effectiveness reviews rarely measure safety, and where safety was examined, adverse events including new-onset suicidal tendency were recorded. Governance should mandate crisis detection and referral, close classification gaps between medical device and general-purpose AI regulation, and require continuous independent auditing.

KEYWORDS

Mental Health Chatbots, Suicide Risk Detection, Large Language Models, Digital Health Safety, Artificial Intelligence Governance

CITATION

Wrzosek, K., Udrycka, K., Kuchta, K., Poleszak, J., Leśniak, M., & Kowalewski, P. (2026). Suicide and crisis risk detection in mental health chatbots: Technical approaches, safety evidence, and governance challenges. *International Journal of Innovative Technologies in Social Science*, 3(51). https://doi.org/10.31435/ijitss.3(51).2026.6484

Introduction

Suicide remains one of the most consequential and least tractable outcomes in mental health care. More than 700,000 people die by suicide each year, and suicide is the second leading cause of death among people aged 15 to 29; although global rates declined by 36% between 2000 and 2019, this decline was not uniform, and rates increased in both South and North America (Pichowicz et al., 2025). The capacity to respond to this burden is constrained by structural limits on access, since although face-to-face psychotherapy and pharmacological treatment are effective, availability is restricted by workforce shortages, stigma and geographical barriers (Huang et al., 2026), and by a documented shortfall in investment in mental health services that intensified after the COVID-19 pandemic (Khawaja & Bélisle-Pipon, 2023).

Conversational agents entered this gap first as rule-based and narrowly scripted systems, and more recently as products built on large language models (LLMs). The evidence base for their therapeutic value has grown quickly and is now moderately large. Meta-analyses report small to moderate reductions in depressive and anxiety symptoms relative to control conditions, with pooled effect sizes of Hedges' g 0.64 for depression and 0.70 for distress in one synthesis of 15 randomised trials (Li et al., 2023), g 0.29 for depressive symptoms across 32 trials involving 6,089 participants (He et al., 2023), and g 0.31 for depression and 0.28 for anxiety across 39 trials (Sohn et al., 2026). The first randomised controlled trial of a fine-tuned generative AI chatbot reported large effects against a waitlist control, with Cohen's d values between 0.627 and 0.903 across depression, anxiety and eating disorder risk (Heinz et al., 2025).

Population exposure has outpaced this evidence base. A nationally representative survey of 1,058 United States youths aged 12 to 21 found that 13.1% had sought advice from generative AI when feeling sad, angry or nervous, rising to 22.2% among those aged 18 to 21; among users, 65.5% did so at least monthly and 92.7% considered the advice helpful (McBain et al., 2025c). Broader usage figures are higher still: 64% of United

States teenagers aged 13 to 17 report using AI chatbots, one in three of them daily, and in the United Kingdom one in four teenagers reports using AI chatbots specifically for mental health support (Cali & van Kolfschooten, 2026). The systems receiving these disclosures are, for the most part, general-purpose assistants or wellness applications rather than regulated clinical products (Heinz et al., 2025).

This mismatch creates a safety problem that is analytically distinct from the question of therapeutic efficacy. A chatbot that reduces mean depression scores across a trial population may still fail catastrophically in the small number of conversations where a user discloses suicidal intent. Evidence that this failure mode is real is now substantial. In a systematic evaluation of 29 commercially available AI chatbot applications using prompts derived from the Columbia-Suicide Severity Rating Scale, no agent met the criteria for an adequate response, and 48.28% of responses were judged inadequate, with recurrent failures to supply emergency contact information (Pichowicz et al., 2025). Qualitative analysis of more than 6,000 user reviews of ten mental health chatbot applications reached a convergent conclusion, namely that even recently released products lack the understanding required to identify a crisis correctly, while simultaneously fostering attachment that displaces contact with friends and family (Haque & Rubya, 2023).

Detecting suicide and crisis risk from text is not itself a new problem. Natural language processing has been applied to mental illness detection for over a decade, with 399 studies identified in one narrative review of 10,467 records (Zhang et al., 2022), and clinical applications have achieved high discrimination when applied retrospectively to electronic health records (Bejan et al., 2022). What is new is the requirement to perform this detection inside a live conversation, in real time, on every turn, in a product that is simultaneously trying to maintain a supportive relationship with the user. This setting changes the problem in at least three respects. First, there is no definitive ground truth for risk, and expert clinicians disagree substantially when labelling the same material (Weber et al., 2026). Second, the costs of the two error types are radically asymmetric and neither is negligible, since a missed disclosure may precede a death while an excess of false alarms degrades the therapeutic experience that motivates use in the first place. Third, the decision about where to set the detection threshold is not a purely technical optimisation, because it determines how often a system will interrupt a conversation, escalate to a human, or refuse to respond.

The regulatory position compounds these difficulties. Emerging regulation of general-purpose AI systems, including the European Union Artificial Intelligence Act, does not classify chatbots as high-risk technologies, so the most stringent obligations do not apply and oversight relies largely on voluntary safeguards implemented by developers, with major platforms differing in when conversations should be terminated, when users should be directed to mental health resources and whether emergency services should be contacted (Cali & van Kolfschooten, 2026). Legal analysis of the interaction between the AI Act and the Medical Device Regulation identifies further lacunae, in which certain AI-specific risks and certain typical medical device risks may remain unregulated (Hauglid & Mahler, 2023).

This narrative review synthesises the literature at the intersection of these three strands. It addresses four questions. What technical approaches have been developed for detecting suicide and crisis risk in conversational and clinical text, and what performance do they achieve? What does the empirical evidence show about the safety behaviour of deployed chatbots and general-purpose LLMs when confronted with suicidal content? What do efficacy trials of mental health chatbots actually measure with respect to safety, and what do they omit? And what governance responses have been proposed, and how do they map onto the technical realities of detection?

Methodology

This study is a narrative review. A narrative rather than systematic design was chosen because the relevant literature is methodologically heterogeneous to a degree that precludes meaningful pooling. It spans retrospective prediction studies on electronic health records, prospective deployment studies in telehealth networks, vignette-based evaluations of commercial LLMs, randomised controlled trials of therapeutic chatbots, meta-analyses of those trials, qualitative analyses of user reviews, ethical analyses, and legal scholarship on European Union product safety regulation. No single effect measure or risk-of-bias instrument applies across these designs. The purpose of the review is accordingly interpretive and integrative rather than estimative: it seeks to characterise the state of technical capability, to establish what safety evidence exists, and to identify where the governance discussion and the technical literature fail to meet.

Records were identified through searches of PubMed, Scopus and Web of Science, conducted in July 2026, combining terms relating to conversational agents (chatbot, conversational agent, virtual agent, large language model, generative artificial intelligence) with terms relating to the outcome domain (suicide, suicidal ideation, self-harm, crisis, risk detection, mental health) and with terms relating to governance (regulation, governance, ethics, explainability). Reference lists of included reviews were checked for additional records.

Records were eligible if they reported empirical data, systematic synthesis, or scholarly analysis bearing on at least one of the four review questions, and if they were published in English. Both peer-reviewed articles and one preprint were included; the preprint is identified as such throughout, and its findings are reported with explicit reference to its unreviewed status. Records were excluded if they concerned conversational agents outside the mental health domain, if they addressed suicide risk without reference to computational methods, or if they duplicated a source already included.

Forty-three records were retained, comprising 42 peer-reviewed journal articles and one preprint, published between 2021 and 2026. The distribution by year was 2 records from 2021, 6 from 2022, 7 from 2023, 11 from 2024, 12 from 2025, and 5 from 2026, reflecting the rapid acceleration of the field following the public release of general-purpose LLM chatbots. The corpus draws predominantly from digital health and medical informatics journals, with the largest contributions from npj Digital Medicine (9 records) and the Journal of Medical Internet Research (6 records), together with records from Nature Medicine, JAMA Network Open, NEJM AI, Psychiatric Services, BMJ Mental Health, Clinical Psychology Review, the European Journal of Public Health, and the Oslo Law Review.

Records were read in full and coded thematically against the four review questions. Where a record contributed to more than one theme, it was coded to each. Quantitative performance figures were extracted verbatim from the source articles, together with the sample and data source to which they apply, and are reported in Tables 1 and 2 with the corresponding principal limitation identified by the original authors. Because the review is narrative, no formal risk-of-bias assessment or certainty rating was applied; instead, the limitations reported by the primary authors are carried forward explicitly into the synthesis, and conclusions are qualified accordingly. Findings from the single preprint are never used as the sole support for a substantive claim.

Results

The first observation the corpus supports is that exposure is already substantial and is concentrated in the age groups at greatest risk. Beyond the survey figures reported above (McBain et al., 2025c; Cali & van Kolfschooten, 2026), the pattern of use matters as much as its prevalence. A survey of 1,006 student users of the intelligent social agent Replika found that respondents were lonelier than typical student populations while still perceiving high social support, that they used the agent in overlapping ways as friend, therapist and intellectual mirror, and that they held simultaneously conflicting beliefs about its nature, describing it as a machine, an intelligence and a human. Critically for the present review, 3% reported that the agent had halted their suicidal ideation (Maples et al., 2024).

This finding is double-edged and illustrates why the safety question cannot be settled by aggregate outcome data. The same relational qualities that make a conversational agent capable of interrupting suicidal ideation also generate what has been termed therapeutic misconception, in which users underestimate the restrictions of the technology and overestimate its capacity to provide genuine therapeutic support, arising through inaccurate marketing, formation of a digital therapeutic alliance, receipt of harmful advice due to

design and algorithmic bias, and the inability of chatbots to foster patient autonomy (Khawaja & Bélisle-Pipon, 2023). Adolescents in particular may experience an empathy gap, a mismatch between apparent empathy and an absence of genuine understanding, while still treating the system as a credible source of advice, with the consequence that emotionally persuasive but inaccurate or unsafe responses may be acted upon (Cali & van Kolfschooten, 2026).

Table 1 summarises the eleven studies in the corpus that report technical performance for suicide or crisis risk detection. Four distinct research settings are represented, and performance is not comparable across them.

Table 1. Technical approaches to suicide and crisis risk detection: data sources, methods, and reported performance

Study	Data source and sample	Technical approach	Reported performance	Principal limitation
Bejan et al. (2022)	Over 200 million clinical notes, single academic health system	Information retrieval with word2vec query expansion and weakly supervised labelling	AUROC 98.6 (95% CI 97.1–99.5) for suicidal ideation and 97.3 (95% CI 95.2–98.7) for suicide attempt among the top 200 retrieved patients; best results when NLP was combined with diagnostic codes and negated expressions were handled explicitly	Single-institution corpus; validation restricted to top-ranked retrieved records
Bayramli et al. (2022)	Electronic health records, structured and unstructured	Naive Bayes and Random Forest classifiers; framework for identifying structured–unstructured feature interactions	Random Forest using both data types reached AUC 0.903 versus 0.887 for structured data alone ($p<0.001$); Naive Bayes showed no comparable gain (0.743 vs 0.742, $p=0.668$)	Retrospective; benefit confined to models able to represent feature interactions
Swaminathan et al. (2023)	721 training messages; retrospective test set N=481; prospective test set N=102,471 in a national telehealth network	Two-stage pipeline: keyword filtering followed by logistic regression	Retrospective AUC 0.82, sensitivity 0.99, PPV 0.35; prospective AUC 0.98, sensitivity 0.98, PPV 0.66; median time from message receipt to specialist triage fell from about nine hours to 8–13 minutes	Single provider network; deliberately low-precision operating point requiring human triage capacity
Grimland et al. (2024)	17,654 crisis hotline chat sessions	SR-BERT trained with a lexicon of theory-derived suicide risk factors	ROC-AUC 92.1%, recall 78.3%, precision 68.9%, F1 73.3%; transformer models outperformed lexicon and Doc2Vec baselines on every metric	Sessions labelled by trained non-professionals; lexicon cannot exhaust theoretical constructs

Study	Data source and sample	Technical approach	Reported performance	Principal limitation
Grimland et al. (2025)	17,564 crisis chat sessions collected 2017–2021	Explainable NLP using a 20-construct psychological lexicon with logistic regression	Previous attempts and hopelessness were the strongest predictors overall; loneliness predicted risk in women and only in adults aged 41 and above; thwarted belongingness strengthened with age; bullying and LGBTQ identity were inversely associated in specific subgroups	Risk operationalised as explicit ideation only; associations are not causal
Thomas et al. (2025)	1,348 helpline encounters with users aged 14–25	Transformer with pretrained weights and LSTM layers; SHAP-based post hoc explanation	Macroaveraged AUC-ROC 0.89 (95% CI 0.81–0.91) and accuracy 0.79; 0.96 for non-suicidal sessions, 0.85 for ideation and 0.87 for advanced suicidal engagement; Brier Skill Score 44% above the word-vector baseline	Modest dataset from one helpline; labels derived from composite C-SSRS scores
Hornstein et al. (2024)	Approximately 19,000 children and young adults, around 800,000 messages, 24/7 crisis service	XGBoost on conversation words with repeated cross-validation and Bayesian hyperparameter search; SHAP explanation	AUROC 0.68 ($p<0.01$) on 3,942 unseen consultations for predicting repeat contact; terms related to self-harm and suicidal thoughts among the strongest features	Predicts re-contact rather than risk itself; discrimination modest
Lho et al. (2025)	1,064 psychiatric patients aged 18–39, 52,627 sentence completion responses	GPT-4o, Gemini 1.0 Pro and GPT-3.5 in zero- and few-shot settings; text-embedding models with gradient boosting	Zero-shot GPT-4o reached AUROC 0.720 for depression and 0.731 for suicide risk; the embedding model with XGBoost reached 0.841 and 0.724; self-concept narratives were most informative	Single centre; projective test format unlike spontaneous chatbot dialogue
Lee et al. (2024)	460 telemental health patients (140 reporting ideation, 120 later reporting ideation with a plan, 200 controls)	GPT-4 compared with six senior clinicians blinded to outcomes, using intake free text	At intake, clinicians achieved higher precision (0.70 vs 0.60) but lower sensitivity (0.53 vs 0.62) than GPT-4; for future ideation with a plan, adding attempt history raised GPT-4 sensitivity to 0.74 with precision 0.48	Intake text only; retrospective; no audit for demographic bias
McCoy & Perlis (2025)	458,053 hospitalised adults; 1,995 deaths by suicide or accident matched 5:1 to controls	Llama-DeepSeek-R1 8B reasoning model applied to discharge notes on consumer-grade hardware	Hazard ratio 4.65 (3.58–6.04); c-statistic 0.64 (0.63–0.65), modestly below larger commercial models, with inspectable chain-of-thought topics	Discrimination modest; retrospective cohort design

Study	Data source and sample	Technical approach	Reported performance	Principal limitation
Weber et al. (2026), preprint	200 clinician-labelled conversation segments drawn from a deployed mental health chatbot; 3,608 interactions screened	One base LLM with five prompt-defined sensitivity thresholds, no task-specific training; five independent expert annotators	False-negative rate fell monotonically from 87% to 0% as the threshold rose; recall reached 98.9% and 100% in the high and extreme variants; mean latency under one second; errors clustered on segments where clinicians themselves disagreed; 92 of 3,608 interactions (2.55%) were ultimately labelled high risk	Preprint without peer review; small dataset with deliberate oversampling of high-risk cases (46%)

The highest reported discrimination in the corpus comes from clinical record mining rather than from conversational settings. Applying information retrieval methods with word2vec query expansion and weak supervision across more than 200 million notes from a single health system, Bejan et al. (2022) reported an AUROC of 98.6 for suicidal ideation and 97.3 for suicide attempt among the top 200 retrieved patients, with the best case extraction achieved by combining NLP with diagnostic codes and by explicitly accounting for negated suicide expressions in notes. The motivation for this work is instructive: diagnostic codes substantially under-ascertain suicide-related outcomes, which undermines estimates of rates, resource allocation and risk assessment alike.

Bayramli et al. (2022) addressed a related question, namely whether unstructured text adds predictive value over structured EHR fields. Their finding was conditional. A Random Forest model trained on both data types reached an AUC of 0.903 against 0.887 for structured data alone ($p<0.001$), whereas a Naive Bayes model showed no equivalent gain (0.743 versus 0.742, $p=0.668$). The authors attributed the difference to the capacity of the Random Forest to capture interactions between the two data types, and proposed a framework for identifying the specific structured-unstructured feature pairs whose interactions differ between cases and non-cases, finding that such pairs tend to capture heterogeneous pairs of general concepts rather than homogeneous pairs of specific concepts.

These results establish an important baseline but transfer poorly to the chatbot setting. Both are retrospective, both operate on clinician-authored text rather than user-authored conversation, and neither has to make a decision in the time available before the next conversational turn.

Crisis helplines and telehealth messaging platforms provide the closest available analogue to the chatbot setting, because the text is user-authored, emotionally charged, and produced in real time. Four studies in the corpus operate in this setting, and their results converge on a consistent picture.

Swaminathan et al. (2023) built and prospectively deployed a two-stage system, keyword filtering followed by logistic regression, in a national telehealth provider network. Trained on 721 messages of which 32% represented potential crises, the model achieved an AUC of 0.82 with sensitivity 0.99 and positive predictive value 0.35 on a retrospective test set of 481 messages, and an AUC of 0.98 with sensitivity 0.98 and PPV 0.66 on a prospective test set of 102,471 messages. The operationally decisive result was not the discrimination metric but the workflow effect: the daily median time from message receipt to crisis specialist triage fell from about nine hours before deployment to between eight and thirteen minutes afterwards. The authors framed their contribution explicitly as human-machine collaboration, concluding that with appropriate training humans can leverage such systems to facilitate triage rather than to replace it.

Two studies from the same research programme analysed a corpus of approximately 17,600 crisis chat sessions from a digital helpline. Grimland et al. (2024) trained SR-BERT, a transformer incorporating a lexicon of language representations of key risk factors derived from major suicide theories, and reported ROC-AUC 92.1%, recall 78.3%, precision 68.9% and F1 73.3%, outperforming a Doc2Vec baseline (ROC-AUC 64.7%) and a lexicon-plus-gradient-boosting model (76.5%). Logistic regression identified hopelessness, history of suicide, self-harm and thwarted belongingness as the most prominent theory-driven correlates. Grimland et al. (2025) extended this analysis to gender and age strata using a lexicon of 20 psychological constructs, finding that previous attempts and hopelessness were the strongest predictors across all groups, that loneliness was a consistent predictor among women but predicted risk only among adults aged 41 and over, that thwarted belongingness strengthened with age, and that prior attempts were strongest in adolescents while hopelessness and self-harm peaked in young adults. Several variables conventionally treated as risk markers, including

bullying, cyberbullying, LGBTQ identity and perfectionism, were inversely associated with risk in specific subgroups.

Thomas et al. (2025) worked with a smaller but more finely labelled dataset of 1,348 encounters from a German crisis helpline serving users aged 14 to 25, with outcomes defined by composite Columbia-Suicide Severity Rating Scale scores. A transformer with pretrained weights and long short-term memory layers achieved a macroaveraged AUC-ROC of 0.89 and accuracy of 0.79, decomposing into 0.96 for non-suicidal sessions, 0.85 for suicidal ideation and 0.87 for advanced suicidal engagement, and improving on a word-vector baseline by 44% on the Brier Skill Score. Shapley Additive Explanations identified self-reference, negation, expressions of low self-esteem and absolutist language as predictive features.

Hornstein et al. (2024) illustrate the limits of the approach when the target is shifted. Using approximately 800,000 messages from around 19,000 children and young adults contacting a 24/7 German crisis service, an XGBoost classifier predicted whether a chatter would contact the service again with an AUROC of 0.68 on 3,942 unseen consultations. Words indicating younger age or female gender and terms related to self-harm and suicidal thoughts were associated with a higher chance of recontacting. The modest discrimination for a proxy outcome, in a dataset far larger than those used for direct risk classification, indicates that performance in this domain is bounded by the predictability of the outcome rather than by data volume.

Four studies evaluate LLMs directly as detection instruments, and their results are markedly more modest than the crisis-chat classifiers.

Lho et al. (2025) analysed sentence completion test narratives from 1,064 psychiatric patients comprising 52,627 completed responses. GPT-4o in zero-shot mode achieved an AUROC of 0.720 for clinically significant depression and 0.731 for high suicide risk using self-concept narratives, with few-shot prompting improving depression detection to 0.754. A text-embedding model paired with extreme gradient boosting outperformed the generative models, reaching 0.841 for depression and 0.724 for suicide risk. Self-concept narratives were the most informative input across all models.

Lee et al. (2024) conducted the most direct comparison with human expertise in the corpus, evaluating GPT-4 against six senior clinicians, three psychologists and three psychiatrists, who were blinded to the subsequent course of treatment. Using chief complaint text alone to identify suicidal ideation with a plan at intake, clinicians achieved higher average precision (0.70 versus 0.60) but lower sensitivity (0.53 versus 0.62) than GPT-4. Adding suicide attempt history raised clinician sensitivity to 0.59 and precision to 0.77, while raising GPT-4 sensitivity to 0.59 but reducing its precision to 0.54. For the harder task of predicting future ideation with a plan, performance fell for both, and with attempt history included GPT-4 reached sensitivity 0.74 at precision 0.48. The authors concluded that GPT-4 approached trained clinician performance on some metrics with a simple prompt design, while stipulating that safety checks for bias must precede any clinical pilot.

McCoy and Perlis (2025) tested whether a reasoning model small enough to run on consumer-grade hardware could approximate larger commercial systems. Applying Llama-DeepSeek-R1 8B to discharge notes from a cohort of 458,053 hospitalised adults, of whom 1,995 died by suicide or accident and were matched 5:1 with controls, they obtained a hazard ratio of 4.65 (3.58 to 6.04) and a c-statistic of 0.64 (0.63 to 0.65), modestly poorer than a larger commercial model. The significance of the study lies less in the discrimination achieved than in the combination of local deployment, which addresses data security, with an inspectable chain of thought, which permits the topics associated with correct and incorrect predictions to be examined directly.

The study closest to the review's central question is a preprint and is reported here with that qualification. Weber et al. (2026) curated 200 conversation segments drawn from a deployed mental health chatbot and had five clinical experts independently annotate each for suicide and crisis risk. Using a single base LLM, they implemented five prompt-defined detection variants with systematically increasing sensitivity thresholds and no task-specific training or fine-tuning. As sensitivity increased, the false-negative rate fell monotonically from 87% to 0%, with the high and extreme variants achieving recall of 98.9% and 100% respectively, at a mean latency below one second. Two secondary findings carry more weight than the headline recall figures. First, model errors aligned closely with the cases on which the clinicians themselves disagreed, which the authors interpret as evidence that misclassification predominantly reflects irreducible uncertainty rather than model failure. Second, the base rate was low: considering the full corpus of 3,608 user-chatbot interactions, only 92 cases (2.55%) were ultimately labelled high risk by clinicians, despite 200 (5.5%) having been flagged by the operational system. The evaluation dataset deliberately oversampled high-risk cases, which constituted 46% of the labelled sample, introducing a prevalence bias that the authors acknowledge.

Explanation methods recur across the detection studies, with Shapley Additive Explanations used by Thomas et al. (2025) and Hornstein et al. (2024), theory-driven lexicons providing interpretable features in both Grimland studies, and chain-of-thought inspection serving the same function for McCoy and Perlis (2025). The corpus also contains dedicated conceptual work on what explanation should mean in this setting, and it is critical of loose usage.

Joyce et al. (2023) observe that the literature on AI and machine learning in mental health and psychiatry lacks consensus on what explainability means, and propose replacing it with understandability, defined as a function of transparency and interpretability, in the Transparency and Interpretability For Understandability framework. They argue that the need for understandability is heightened in psychiatry specifically, because the data describing syndromes, outcomes, disorders and signs possess probabilistic rather than deterministic relationships to one another, as do the tentative aetiologies and multifactorial determinants of disorders. Huang et al. (2024) reach a compatible conclusion from a scoping review of explainable and interpretable deep learning in healthcare NLP, distinguishing explainable from interpretable AI, finding attention mechanisms to be the most prevalent emerging interpretable technique, and identifying as major challenges the failure of most work to explore global modelling processes, the absence of best practices, and the lack of systematic evaluation and benchmarks. Zhang et al. (2022), reviewing 399 studies of NLP for mental illness detection, similarly recommend the development of interpretable models as a priority direction, having found that deep learning methods receive more attention and perform better than traditional machine learning approaches.

Table 2 summarises the fourteen records that bear directly on safety. They divide into evaluations of how systems actually behave when confronted with suicidal content, and evidence about what the efficacy literature does and does not measure.

Table 2. Safety evidence on deployed chatbots and general-purpose large language models

Study	Object of evaluation	Method	Safety-relevant finding
Pichowicz et al. (2025)	29 commercially available AI chatbot applications	Standardised escalating prompts derived from the Columbia-Suicide Severity Rating Scale	No agent met the criteria for an adequate response; 51.72% met relaxed criteria for a marginal response and 48.28% were inadequate. Recurrent failures were absence of emergency contact information and lack of contextual understanding
Haque & Rubya (2023)	10 popular mental health chatbot applications	Qualitative analysis of 3,621 Google Play and 2,624 App Store reviews	Users reported becoming over-attached and preferring the chatbot to friends and family; even recently released chatbots lacked the understanding needed to identify a crisis correctly
McBain et al. (2025a)	ChatGPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro	Revised Suicidal Ideation Response Inventory across 24 scenarios, benchmarked against expert suicidologists	All three models judged clinician responses as more appropriate than experts did, with mean item differences of 0.86, 0.61 and 0.73; outlier ratings occurred in 19%, 11% and 36% of items; final scores of 45.7, 36.7 and 54.5 corresponded to performance ranging from untrained school staff to trained professionals
McBain et al. (2025b)	ChatGPT, Claude, Gemini	30 suicide-related queries stratified by 13 clinical experts into five risk levels; 9,000 responses	ChatGPT and Claude responded directly to 100% of very low-risk and 0% of very high-risk queries, but did not meaningfully differentiate the three intermediate levels; Claude was more likely to answer directly (aOR 2.01) and Gemini less likely (aOR 0.91) than ChatGPT

Study	Object of evaluation	Method	Safety-relevant finding
Lauderdale et al. (2025)	ChatGPT-3.5, ChatGPT-4o, Google Gemini	Four standardised veteran vignettes rated against the Veterans Health Administration risk stratification table and compared with mental health care providers	Ratings diverged at the extremes, with the most acute case rated less acute and the least acute case rated more acute; ChatGPT-4o recommended hospitalisation for every veteran; treatment planning was predicted by chronic rather than acute risk ratings
Campbell et al. (2025)	Nine general-purpose generative chatbots	Content analysis of responses to suicide-related questions in two phases one year apart, with linguistic analysis	Depth and accuracy improved between phases, with greater emphasis on the 988 crisis line; the authors nonetheless conclude that referral to trained professionals remains necessary and that outputs require periodic re-checking
Hager et al. (2024)	State-of-the-art LLMs in clinical decision-making	2,400 real patient cases from an intensive care database within a simulated clinical workflow	Models diagnosed significantly worse than physicians, followed neither diagnostic nor treatment guidelines, could not interpret laboratory results, and were sensitive to the quantity and ordering of information, leading the authors to conclude that LLMs are not ready for autonomous clinical decision-making
Maples et al. (2024)	Replika, an intelligent social agent	Survey of 1,006 student users	Respondents were lonelier than typical student populations; 3% reported that the agent had halted their suicidal ideation; users simultaneously described the agent as a machine, an intelligence and a human
McBain et al. (2025c)	Population-level exposure	Nationally representative survey of 1,058 US youths aged 12–21	13.1% had used generative AI for mental health advice, rising to 22.2% among those aged 18–21; 65.5% of users did so monthly or more often and 92.7% found the advice helpful
Heinz et al. (2025)	Therabot, a fine-tuned generative AI chatbot	Randomised controlled trial, 210 adults with depression, anxiety or high risk for feeding and eating disorders	Large symptom reductions (d 0.845–0.903 for depression, 0.794–0.840 for anxiety, 0.627–0.819 for eating disorder risk), but active suicidality was an exclusion criterion, a separate crisis classification model and emergency module were built in, and every response was reviewed by trained clinicians after transmission, with direct contact when safety concerns arose
Li et al. (2023)	AI-based conversational agents for mental health	Systematic review of 35 studies drawn from 7,834 records	Despite the growing significance of safety concerns, only 15 of the 35 studies incorporated any safety assessment or protection measure, such as access to human experts, onboarding processes, assessment of adverse events, or automatic crisis and harm identification
Balan et al. (2024)	Automated conversational agents for young people	Scoping review of 25 studies, N=1,707	Safety issues were reported in only two studies; in one, more than half of participants reported at least one negative effect and a serious adverse event occurred in which a participant reported suicidal tendency for the first time after the intervention; in another, four participants (24%) had one suicidal ideation alert, four had three and two had six, and one parent reported that their child had been seen in an emergency department

Study	Object of evaluation	Method	Safety-relevant finding
Sohn et al. (2026)	Chatbots for depressive and anxiety symptoms	Systematic review and meta-analysis of 39 randomised trials searched to October 2025	The authors conclude that safety monitoring and transparent adverse-event reporting remain underdeveloped, and recommend that future trials implement prespecified safety protocols and systematically report adverse events associated with chatbot use
Jabir et al. (2024)	Conversational agent interventions	Systematic review and meta-analysis of 41 randomised trials	Pooled attrition in intervention groups was 21.84% (95% CI 16.74–27.36), rising to 26.59% in studies longer than eight weeks and higher in stand-alone agents without human support

The most direct evidence concerns commercial applications marketed for mental distress. Pichowicz et al. (2025) searched application repositories for products claiming benefit in mental distress and offering an AI-powered chatbot function, identified 29 such agents, and tested each with a standardised set of prompts based on the Columbia-Suicide Severity Rating Scale designed to simulate increasing suicidal risk. None of the tested agents satisfied the initial criteria for an adequate response. Under relaxed criteria, 51.72% produced a marginal response and 48.28% were deemed inadequate, with common errors including inability to provide emergency contact information and a lack of contextual understanding. The authors concluded that these findings raise concerns about deployment in sensitive health contexts without proper clinical validation.

The user-side view is consistent. Haque and Rubya (2023) examined ten popular mental health chatbot applications and analysed 3,621 reviews from the Google Play Store and 2,624 from the Apple App Store. Personalised, humanlike interaction was received positively, but improper responses and unfounded assumptions about users' personalities produced loss of interest. The reviewers found that constant availability leads some users to become overly attached and to prefer the chatbot to interaction with friends and family, and that although a chatbot may in principle offer crisis care at any hour because of its availability, even recently developed chatbots lack the understanding required to identify a crisis properly. The authors warn that excessive reliance on the technology carries risks including isolation and insufficient assistance during crises.

A second group of studies asks not whether systems respond, but whether their judgements match those of qualified clinicians. The results indicate that alignment holds at the extremes of the risk continuum and breaks down in the middle.

McBain et al. (2025b) had thirteen clinical experts categorise 30 hypothetical suicide-related queries into five levels of self-harm risk, then elicited 100 responses per query from each of three chatbots, yielding 9,000 responses coded as direct or indirect. ChatGPT and Claude provided direct responses 100% of the time for very low-risk queries and 0% of the time for very high-risk queries, and Gemini was more variable. Critically, none of the models meaningfully distinguished among intermediate risk levels: relative to very low-risk queries, the odds of a direct response were not statistically different for low, medium or high-risk queries. Claude was more likely than ChatGPT to respond directly (aOR 2.01, 95% CI 1.71 to 2.37) and Gemini less likely (aOR 0.91).

The companion study applied the revised Suicidal Ideation Response Inventory, a training instrument for mental health professionals comprising 24 scenarios each followed by two clinician responses scored from -3 to +3 (McBain et al., 2025a). All three models rated responses as more appropriate than expert suicidologists did, with item-level mean differences of 0.86 for ChatGPT, 0.61 for Claude and 0.73 for Gemini. Outlier ratings occurred for 19% of ChatGPT items, 11% of Claude items and 36% of Gemini items. Final scores placed ChatGPT at roughly the level of master's-level counsellors, Claude above the performance of mental health professionals following suicide intervention skills training, and Gemini at the level of untrained school staff. The consistent direction of the bias is the salient finding: models systematically judged responses to suicidal ideation as more appropriate than experts did.

Lauderdale et al. (2025) tested three models against the Veterans Health Administration risk stratification table using four standardised vignettes. Discrepancies with mental health care providers appeared at both ends, with the most acute case rated as less acute and the least acute case rated as more acute. Variation across models was also substantial: ChatGPT-3.5 gave lower acute risk ratings than ChatGPT-4o and Gemini, while ChatGPT-4o assigned higher chronic risk ratings and recommended hospitalisation for all four veterans.

Treatment planning by the models was predicted by chronic but not acute risk ratings. The authors concluded that continued clinician oversight is essential.

Campbell et al. (2025) offer the corpus's only longitudinal observation of general-purpose systems, conducting a content analysis of chatbot responses to suicide-related questions in two phases one year apart, covering nine systems in total. Depth and accuracy of responses increased between phases, with responses becoming more comprehensive across signs of suicide, lethality, resources and ways to support those in crisis, and with greater emphasis on the 988 crisis line. The authors nonetheless conclude that the importance of seeking help from a trained mental health professional remains, and that generative AI responses to suicide questions should be checked periodically to ensure that suicide prevention best practice is being communicated. The implication for governance is that safety behaviour is a moving target which changes without notice between model versions.

The wider evidence on clinical reasoning provides context for these findings. Hager et al. (2024) constructed a dataset of 2,400 real patient cases across four abdominal pathologies within a framework simulating a realistic clinical environment, and found that state-of-the-art LLMs did not accurately diagnose patients, performed significantly worse than physicians, followed neither diagnostic nor treatment guidelines, could not interpret laboratory results, and were sensitive to both the quantity and the order of information presented. Their conclusion that LLMs are not currently ready for autonomous clinical decision-making applies with additional force in a domain where the relevant judgement is less codified than abdominal diagnosis.

The therapeutic evidence base for chatbots is substantial but was not designed to answer safety questions, and the corpus documents this gap explicitly.

The most direct statement of the gap comes from the reviews that counted how often safety was addressed at all. Li et al. (2023), synthesising 35 studies of AI-based conversational agents, observed that despite the growing significance of safety concerns in this domain, only 15 of those studies incorporated any safety assessment or protection measure, the measures in question being access to human experts, onboarding processes, assessment of adverse events, and automatic crisis or harm identification. He et al. (2023) note in the same vein that prior work has revealed hazards associated with conversational agent interventions, including misunderstandings that may result in inefficient or even harmful interventions, a lack of crisis warning systems, and a lack of privacy protection. Three years later the position had not materially changed: Sohn et al. (2026) conclude that safety monitoring and transparent adverse-event reporting remain underdeveloped, and place among their recommendations for future trials the implementation of prespecified safety protocols and the systematic reporting of adverse events associated with chatbot use.

Where safety has been examined, the events recorded are not trivial. Balan et al. (2024), in a scoping review of 25 studies involving 1,707 young participants, found that safety issues were reported in only two of them. In one, more than half of the participants reported at least one negative effect of the intervention, and a serious adverse event occurred in which a participant reported suicidal tendency for the first time after the intervention. In the other, four participants (24%) registered one alert for suicidal ideation, four registered three and two registered six, and one parent reported in week 12 that their child had been seen in an emergency department and discharged. These are the only two studies out of twenty-five in which such events could have been detected, which makes the aggregate reassurance offered by the effectiveness literature difficult to interpret.

The effectiveness syntheses themselves are numerous and broadly consistent. Li et al. (2023) screened 7,834 records, included 35 studies and meta-analysed 15 randomised trials, finding significant reductions in depression (g 0.64, 95% CI 0.17 to 1.12) and distress (g 0.70, 95% CI 0.18 to 1.22), with more pronounced effects for agents that were multimodal, generative AI-based, integrated with mobile or instant messaging applications, and targeted at clinical, subclinical and elderly populations. He et al. (2023) pooled 32 trials with 6,089 participants and reported significant short-term effects across nine outcomes with effect sizes between 0.24 and 0.62, while long-term effects were largely not significant, ranging from -0.04 to 0.39; personalisation and empathic response emerged as critical facilitators of efficacy. Sohn et al. (2026), searching to October 2025, analysed 38 trials for depression (n=7,401) and 34 for anxiety (n=7,621) and reported smaller pooled effects of g 0.31 and g 0.28, with larger effects in clinical and subclinical than in non-clinical samples. Villarreal-Zegarra et al. (2024) restricted attention to self-administered interventions based on NLP models, including 21 articles of which 16 entered meta-analysis, and reported standardised mean differences of 0.819 for depressive and 0.272 for anxiety symptoms. Huang et al. (2026), examining eleven trials with 2,220 participants and focusing on interaction features, found that the pooled effect on depressive symptoms was not

statistically significant (SMD -0.46, 95% CI -1.02 to 0.10), while emotional responsiveness, structured feedback strategies and interaction frequency were associated with higher adherence.

Individual trials in the corpus display the same pattern of measuring symptoms rather than safety. Trials of XiaoE in Chinese university students (He et al., 2022), Tess in Argentinian students (Klos et al., 2021), an AI chatbot for Chinese university students (Liu et al., 2022), and the Woebot adaptation for substance use (Prochaska et al., 2021) all report symptom or engagement outcomes. One trial reports a negative finding of direct relevance: Karkosz et al. (2024) tested Fido, a Polish-language CBT chatbot, in 81 participants with subclinical depression or anxiety and did not replicate previous chatbot findings, since symptoms fell in both arms and remained stable at one-month follow-up, with the control group in fact spending more time on its intervention than the chatbot group.

Engagement data qualify the therapeutic picture further. Jabir et al. (2024) meta-analysed attrition across 41 randomised trials and found a pooled intervention-group dropout rate of 21.84% (95% CI 16.74 to 27.36), rising to 26.59% in studies lasting more than eight weeks, with intervention participants more likely to drop out than controls in both short-term and long-term studies, and with higher attrition in stand-alone agents lacking human support. Balan et al. (2024) likewise report that automated conversational agents for young people were acceptable, engaging and usable overall, while noting that most were stand-alone preventive tools and that two studies documented a decline in engagement and adherence over time. Attrition is not itself an adverse event, but in a system whose safety function depends on continued contact, disengagement removes the very channel through which risk would be detected.

The single most instructive record on the efficacy-safety gap is the trial that produced the largest effects. Heinz et al. (2025) randomised 210 adults to a four-week Therabot intervention or a waitlist control and reported differential symptom reductions with Cohen's d between 0.845 and 0.903 for depression, 0.794 and 0.840 for generalised anxiety, and 0.627 and 0.819 for eating disorder risk, together with therapeutic alliance ratings comparable to those given to human therapists. The safety architecture surrounding this result deserves equal attention. Active suicidality was an exclusion criterion. Multiple guard rails were added because of the risks associated with generative AI, including a crisis classification model and an emergency module deployed in response to model detection of high-risk content such as suicidal ideation. All responses from Therabot were supervised by trained clinicians and researchers after transmission; when an inappropriate response occurred, the research team contacted the participant to provide correction, and when a participant raised safety concerns such as suicidal ideation, the team contacted the participant to provide safety guidance and emergency resources. The strongest efficacy evidence in the corpus therefore derives from conditions in which autonomous crisis detection was not the operative safety mechanism.

Broader syntheses of AI performance in mental health provide the outer bound of what should be expected. An umbrella review of 15 systematic reviews drawn from 852 citations reported classifier accuracy in diagnosing mental disorders ranging from 68% to 100%, with one meta-analysed review reporting pooled sensitivity of 92% and specificity of 86% (Abd-Alrazaq et al., 2022). A meta-analysis of 155 studies predicting binary treatment response in emotional disorders found a mean prediction accuracy of 0.76 (95% CI 0.74 to 0.78) and mean AUC of 0.80, with sensitivity 0.73 and specificity 0.75, and reported that studies using more robust cross-validation procedures exhibited higher accuracy (Curtiss & DiPietro, 2025). A meta-analysis of AI for postpartum depression detection across 16 studies reported pooled sensitivity of 69% (95% CI 55 to 81) and accuracy of 79% (95% CI 73 to 85), alongside reported challenges of algorithmic bias, data privacy and implementation barriers (Ruger-Navarrete et al., 2026).

The governance literature in the corpus addresses three questions: whether mental health chatbots fall within existing regulatory categories, what ethical obligations attach to them irrespective of regulatory classification, and what targeted responses would fit their specific features.

On classification, Hauglid and Mahler (2023) analyse the interplay between the proposed EU AI Act and the Medical Device Regulation from a product safety perspective, using two case studies displaying different degrees of generativity. Their central finding is a structural one. Because the AI Act relies on MDR classification for medical AI, systems that fall into MDR class I, or fall outside the scope of the MDR because they are not medical devices, are not covered by the AI Act's high-risk category unless separately listed in Annex III. The consequence is that the emerging framework potentially leaves both certain AI-specific risks and certain typical medical device risks unregulated. They propose two remedies: adding certain medical AI systems to Annex III of the AI Act, using the MDR definition of a medical device as the basis for a targeted inclusion, which would capture a system's degree of generativity as a criterion; or amending the MDR classification rules with a rule for generative AI medical devices capable of providing treatment responses

autonomously, which is more targeted but would not remedy the gap for risky systems that are not defined as medical devices at all.

Cali and van Kolfschooten (2026) approach the same gap from public health rather than product safety, arguing that AI chatbots should be understood as emerging public health technologies, meaning technologies whose scale, design and capacity to influence health-related behaviours, wellbeing and access to care warrant governance based on public health objectives rather than only consumer protection or innovation concerns. They observe that regulation of general-purpose AI systems, including the EU AI Act, does not classify chatbots as high-risk, so that the most stringent obligations do not apply and oversight relies largely on voluntary safeguards, and that regulatory oversight remains fragmented in practice: major platforms differ, according to their terms of use, in when conversations should be terminated, when users should be directed to mental health resources, and whether emergency services should be contacted. They further note that frameworks developed in response to youth social media use are organised by platform category, so that chatbots, not being social media platforms, frequently fall outside their scope. Their proposed responses map directly onto the technical material reviewed above. First, minimum safety standards for youth-facing AI systems, including explicit requirements for crisis detection and appropriate referral to qualified mental health services, supported by clinical evaluation standards analogous to those applied to digital health interventions. Second, governance of relational and persuasive design features, particularly where systems simulate emotional support or encourage prolonged interaction among adolescents. Third, independent auditing mechanisms to evaluate chatbot responses to health-related queries for accuracy, safety and consistency across platforms.

The ethical analyses in the corpus supply the normative content for these mechanisms. Khawaja and Bélisle-Pipon (2023) identify therapeutic misconception as the central risk, arising when users underestimate the restrictions of the technology and overestimate its ability to provide therapeutic support, and trace its pathways through inaccurate marketing, the formation of a digital therapeutic alliance, harmful advice arising from design and algorithmic bias, and the inability of chatbots to foster autonomy. Blease and Rodman (2024) frame generative AI in mental healthcare within the four principles of biomedical ethics and identify the empirical questions that must be answered to inform ethical practice, including, under nonmaleficence, whether use increases self-harm episodes and whether patients and clinicians perceive errors in generative AI outputs, and under justice, whether patients perceive stigmatised language and whether perceptions differ across demographic groups.

Discussion

The performance figures assembled in Table 1 span a wide range, which invites the conclusion that some approaches are far better than others. That reading would be an artefact, because the studies do not predict the same thing. Where the label is derived from clinical documentation, it inherits the under-ascertainment that motivated the work in the first place. Where it is applied to conversations, it comes variously from trained non-professionals, from composite scores on a standardised instrument, and in one case from a proxy chosen for its operational usefulness rather than its equivalence to risk. The finding that model errors concentrate on precisely the segments about which expert annotators disagreed, if it holds in larger samples, indicates that a portion of what is reported as classifier error is not error in any correctable sense but disagreement about an underdetermined judgement. Two things follow. Improvements in discrimination beyond a certain point may register fidelity to one labelling protocol rather than a greater capacity to identify people at risk. And a field that competes on discrimination without reporting label provenance and inter-rater agreement alongside it will keep producing numbers that neither compare nor transfer.

The trade-off between missed disclosures and false alarms is therefore not a problem to be solved but a parameter to be set, and the evidence that a single base model can be moved across almost the entire range of that trade-off by threshold alone makes the point sharply. Near-zero-miss detection is available; a threshold that avoids both error types is not. The choice between them cannot be made on statistical grounds, because the two errors are incommensurable: one may precede a death, the other consumes human capacity and, where the response is to interrupt or refuse, erodes the supportive interaction that brought the user to the system. Rarity compounds this. When the condition occurs in a small fraction of interactions, precision is capped by the base rate irrespective of model quality, which is why positive predictive values across this corpus sit so far below the corresponding sensitivities. The alert volume that a safe threshold generates is thus a design

constraint of the same order as model performance, and it is ultimately a claim about staffing. A telehealth network with employed crisis specialists could absorb that volume and convert it into a substantial reduction in triage delay. A consumer application distributed at population scale has no comparable mechanism, which is one reason why the products tested here appear to operate at thresholds that generate almost no alerts at all.

Read together, these findings relocate the safety problem from the model to the system around it. In every case in this corpus where crisis handling demonstrably worked, the same structure was present: a detection function separated from the conversational function and coupled to a human escalation pathway. The most instructive instance is the trial that produced the strongest efficacy result, which excluded active suicidality at enrolment, ran a separate crisis classifier with a dedicated emergency module, and had clinicians review responses after transmission and contact participants directly when safety concerns arose. This is a strength of the trial, but it bounds what the trial licenses. It demonstrates that a fine-tuned generative chatbot can reduce symptoms under clinical supervision; it does not demonstrate that such a system is safe when deployed without that supervision to a population that has not been screened. The distance between those two propositions is the distance between the trial and the market.

The evidence on general-purpose models reinforces the same conclusion from the other direction. Their alignment with clinical judgement holds at the extremes of the risk continuum and disappears in the intermediate range, and the intermediate range is where risk assessment does its work. Explicit disclosures are lexically unambiguous and are handled by refusal and referral; trivial queries require no special handling. Ambiguous, escalating and partially disclosed presentations are both the most common and the most consequential, and there the evaluated models supply no discrimination, while at the same time judging responses to suicidal ideation more favourably than expert suicidologists do and varying substantially among themselves in the risk ratings and treatment recommendations they produce for identical material. Set against the demonstration that current models fail to follow guidelines and are sensitive to the ordering of information in a clinical domain far better codified than suicide risk assessment, the case for retaining a human decision point in the escalation pathway rests on evidence rather than on caution.

This is why the effectiveness literature cannot bear the weight routinely placed on it. That literature was designed to detect symptom change and it does so; it was not designed to detect harm, and the reviews within it say as much. Where safety was examined, events were found. The inferential error this invites is to read unmeasured harm as absent harm, and it is an error that the growing volume of positive effect estimates makes easier rather than harder to commit. A minimum reporting standard follows directly from the corpus itself: prespecified safety protocols and systematic adverse-event reporting, as one of the reviews recommends; an explicit statement of whether suicidality was an exclusion criterion and what escalation pathway operated during the trial, following the practice of the strongest trial in the field; and treatment of attrition as a safety-relevant quantity rather than a nuisance parameter, since in a system whose protective function depends on continued contact, disengagement removes the channel through which risk would be detected.

Explanation methods occupy an ambiguous position in this picture. They demonstrably produce clinically legible output, and the features they surface are consistent with established theory and vary in ways that match the clinical literature on age and gender. But legibility is not validation, and the conceptual work in this corpus warns against treating the two as equivalent, both in its argument that the field lacks an agreed meaning for explainability and in its finding that systematic evaluation and benchmarks are largely absent. What explanation supplies is not a safeguard in itself but the precondition for independent auditing. That connects directly to the structural problem identified in the governance literature, which no amount of model improvement will resolve: because high-risk designation is tied to medical device classification, systems can escape the most stringent obligations through the seam between the two regimes, and because other frameworks are organised by platform category, chatbots fall outside them as well. The proposed remedies map closely onto the technical findings reviewed here. One finding, however, constrains the form that auditing must take. Safety behaviour in general-purpose systems was observed to improve between two assessments a year apart, with no regulatory event in between, which means it is revised silently as models are updated. Certification at a point in time cannot establish a property that changes between versions, and the implication is continuous post-market surveillance rather than pre-market approval alone.

Several limitations qualify these conclusions. The review is narrative and its conclusions are interpretive rather than estimative; no formal risk-of-bias assessment or certainty rating was applied, no quantitative pooling was performed, and the limitations reported by primary authors have been carried forward but not independently verified. The corpus is restricted to English-language records and to 43 items, and a systematic search with duplicate screening would identify records it does not contain. Three specific constraints deserve

emphasis. The study that speaks most directly to the central question is a preprint that has not undergone peer review, rests on a small labelled sample, and deliberately oversampled high-risk cases, so its findings have been treated as hypothesis-generating and are nowhere the sole support for a claim. The evaluations of commercial systems are snapshots of products that change without notice, so figures attached to named models should be read as historical rather than current. And the meta-analytic evidence carries its own constraints, including indications of possible publication bias, predominant reliance on self-report rather than clinician-rated instruments, and high risk of bias in most included trials owing to the impossibility of maintaining participant blinding.

Conclusions

Suicide and crisis risk detection in mental health chatbots is not, on the present evidence, primarily a modelling problem. Detection from natural language is technically feasible in real time and at high sensitivity, and the studies reviewed here demonstrate discrimination in crisis conversation corpora that is adequate for a screening function. What the evidence does not support is the proposition that this capability, embedded in a conversational product and operating autonomously, constitutes a safeguard.

Three findings sustain that conclusion. Detection accuracy is bounded by an underdetermined ground truth, which means that residual error cannot be engineered away and must instead be absorbed by system design. The choice of detection threshold determines the balance between missed disclosures and alert volume, and is therefore a decision about how much human review a deployment will fund rather than a parameter to be optimised. And where crisis handling has demonstrably worked, from a telehealth triage deployment to the strongest randomised trial in the field, it worked because detection was separated from conversation and coupled to a human escalation pathway, not because a model was accurate enough to act alone.

Against this, general-purpose and commercial systems fail in the ways that matter most, aligning with clinical judgement only at the extremes of the risk continuum, judging responses to suicidal ideation more favourably than expert suicidologists, and, in the case of the commercial applications tested to date, failing to meet the criteria for an adequate response to escalating suicidal risk. Meanwhile the effectiveness literature routinely cited in support of these products was not designed to detect harm, has rarely measured it, and where it has, has found it.

The governance implications are correspondingly specific. Minimum safety standards for youth-facing systems should require crisis detection and referral as a functional obligation, since the capability exists. Classification frameworks should be amended so that systems capable of providing treatment responses autonomously do not escape high-risk designation through gaps between medical device and general-purpose AI regulation. Auditing should be independent, should cover consistency across platforms, and should be continuous rather than point-in-time, because safety behaviour changes silently with model updates. Trials should report prespecified safety protocols, adverse events, exclusion of suicidality, and the escalation pathway in operation, so that readers can distinguish what a supervised study demonstrates from what an unsupervised product would deliver. Until that distinction is routinely reported, the growing evidence of symptom improvement will continue to be read as evidence of safety, which it is not.

Conflicts of Interest: No conflicts of interest to declare.

REFERENCES

1. Abd-Alrazaq, A., Alhuwail, D., Schneider, J., Toro, C. T., Ahmed, A., Alzubaidi, M., Alajlani, M., & Househ, M. (2022). The performance of artificial intelligence-driven technologies in diagnosing mental disorders: An umbrella review. *npj Digital Medicine*, *5*(1), 87. https://doi.org/10.1038/s41746-022-00631-8

2. Balan, R., Dobrean, A., & Poetar, C. R. (2024). Use of automated conversational agents in improving young population mental health: A scoping review. *npj Digital Medicine*, *7*(1), 75. https://doi.org/10.1038/s41746-024-01072-1

3. Bayramli, I., Castro, V., Barak-Corren, Y., Madsen, E. M., Nock, M. K., Smoller, J. W., & Reis, B. Y. (2022). Predictive structured–unstructured interactions in EHR models: A case study of suicide prediction. *npj Digital Medicine*, *5*(1), 15. https://doi.org/10.1038/s41746-022-00558-0

4. Bejan, C. A., Ripperger, M., Wilimitis, D., Ahmed, R., Kang, J., Robinson, K., Morley, T. J., Ruderfer, D. M., & Walsh, C. G. (2022). Improving ascertainment of suicidal ideation and suicide attempt with natural language processing. *Scientific Reports*, *12*(1), 15146. https://doi.org/10.1038/s41598-022-19358-3

5. Blease, C., & Rodman, A. (2024). Generative artificial intelligence in mental healthcare: An ethical evaluation. *Current Treatment Options in Psychiatry*, *12*(1), 5. https://doi.org/10.1007/s40501-024-00340-x

6. Cali, A., & van Kolfschooten, H. (2026). Beyond social media: Public health governance of artificial intelligence-based chatbots used by adolescents. *European Journal of Public Health*, *36*(4), ckag079. https://doi.org/10.1093/eurpub/ckag079

7. Campbell, L. O., Babb, K., Lambie, G. W., & Hayes, B. G. (2025). An examination of generative AI response to suicide inquires: Content analysis. *JMIR Mental Health*, *12*, e73623. https://doi.org/10.2196/73623

8. Curtiss, J., & DiPietro, C. (2025). Machine learning in the prediction of treatment response for emotional disorders: A systematic review and meta-analysis. *Clinical Psychology Review*, *120*, 102593. https://doi.org/10.1016/j.cpr.2025.102593

9. Grimland, M., Benatov, J., Yeshayahu, H., Izmaylov, D., Segal, A., Gal, K., & Levi-Belz, Y. (2024). Predicting suicide risk in real-time crisis hotline chats integrating machine learning with psychological factors: Exploring the black box. *Suicide and Life-Threatening Behavior*, *54*(3), 416–424. https://doi.org/10.1111/sltb.13056

10. Grimland, M., Liberman, M., Yeshayahu, H., Benatov, J., Munz, N., Segal, A., Ben Dayan, L., Shenfeld, I., Gal, K., & Levi-Belz, Y. (2025). Explainable AI for suicide risk detection: Gender- and age-specific patterns from real-time crisis chats. *Frontiers in Medicine*, *12*, 1703755. https://doi.org/10.3389/fmed.2025.1703755

11. Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, *30*(9), 2613–2622. https://doi.org/10.1038/s41591-024-03097-1

12. Haque, M. D. R., & Rubya, S. (2023). An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. *JMIR mHealth and uHealth*, *11*, e44838. https://doi.org/10.2196/44838

13. Hauglid, M. K., & Mahler, T. (2023). Doctor chatbot: The EU's regulatory prescription for generative medical AI. *Oslo Law Review*, *10*(1), 1–23. https://doi.org/10.18261/olr.10.1.1

14. He, Y., Yang, L., Zhu, X., Wu, B., Zhang, S., Qian, C., & Tian, T. (2022). Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: Single-blind, three-arm randomized controlled trial. *Journal of Medical Internet Research*, *24*(11), e40719. https://doi.org/10.2196/40719

15. He, Y., Yang, L., Qian, C., Li, T., Su, Z., Zhang, Q., & Hou, X. (2023). Conversational agent interventions for mental health problems: Systematic review and meta-analysis of randomized controlled trials. *Journal of Medical Internet Research*, *25*, e43862. https://doi.org/10.2196/43862

16. Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*, *2*(4). https://doi.org/10.1056/AIoa2400802

17. Hornstein, S., Scharfenberger, J., Lueken, U., Wundrack, R., & Hilbert, K. (2024). Predicting recurrent chat contact in a psychological intervention for the youth using natural language processing. *npj Digital Medicine*, *7*(1), 132. https://doi.org/10.1038/s41746-024-01121-9

18. Huang, G., Li, Y., Jameel, S., Long, Y., & Papanastasiou, G. (2024). From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? *Computational and Structural Biotechnology Journal*, *24*, 362–373. https://doi.org/10.1016/j.csbj.2024.05.004

19. Huang, T., Li, S., Wang, Y., & Liu, W. (2026). Therapeutic interaction features of AI chatbots in depression interventions: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *28*, e88697. https://doi.org/10.2196/88697

20. Jabir, A. I., Lin, X., Martinengo, L., Sharp, G., Theng, Y.-L., & Tudor Car, L. (2024). Attrition in conversational agent–delivered mental health interventions: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *26*, e48168. https://doi.org/10.2196/48168

21. Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine*, *6*(1), 6. https://doi.org/10.1038/s41746-023-00751-9

22. Karkosz, S., Szymański, R., Sanna, K., & Michałowski, J. (2024). Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: Randomized controlled trial. *JMIR Formative Research*, *8*, e47960. https://doi.org/10.2196/47960

23. Khawaja, Z., & Bélisle-Pipon, J.-C. (2023). Your robot therapist is not your therapist: Understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, *5*, 1278186. https://doi.org/10.3389/fdgth.2023.1278186

24. Klos, M. C., Escoredo, M., Joerin, A., Lemos, V. N., Rauws, M., & Bunge, E. L. (2021). Artificial intelligence–based chatbot for anxiety and depression in university students: Pilot randomized controlled trial. *JMIR Formative Research*, *5*(8), e20678. https://doi.org/10.2196/20678

25. Lauderdale, S. A., Schmitt, R., Wuckovich, B., Dalal, N., Desai, H., & Tomlinson, S. (2025). Effectiveness of generative AI-large language models' recognition of veteran suicide risk: A comparison with human mental health providers using a risk stratification model. *Frontiers in Psychiatry*, *16*, 1544951. https://doi.org/10.3389/fpsyt.2025.1544951

26. Lee, C., Mohebbi, M., O'Callaghan, E., & Winsberg, M. (2024). Large language models versus expert clinicians in crisis prediction among telemental health patients: Comparative study. *JMIR Mental Health*, *11*, e58129. https://doi.org/10.2196/58129

27. Lho, S. K., Park, S.-C., Lee, H., Oh, D. Y., Kim, H., Jang, S., Jung, H. Y., Yoo, S. Y., Park, S. M., & Lee, J.-Y. (2025). Large language models and text embeddings for detecting depression and suicide in patient narratives. *JAMA Network Open*, *8*(5), e2511922. https://doi.org/10.1001/jamanetworkopen.2025.11922

28. Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C. (2023). Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digital Medicine*, *6*(1), 236. https://doi.org/10.1038/s41746-023-00979-5

29. Liu, H., Peng, H., Song, X., Xu, C., & Zhang, M. (2022). Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interventions*, *27*, 100495. https://doi.org/10.1016/j.invent.2022.100495

30. Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Loneliness and suicide mitigation for students using GPT3-enabled chatbots. *npj Mental Health Research*, *3*(1), 4. https://doi.org/10.1038/s44184-023-00047-6

31. McBain, R. K., Cantor, J. H., Zhang, L. A., Baker, O., Zhang, F., Halbisen, A., Kofner, A., Breslau, J., Stein, B., Mehrotra, A., & Yu, H. (2025a). Competency of large language models in evaluating appropriate responses to suicidal ideation: Comparative study. *Journal of Medical Internet Research*, *27*, e67891. https://doi.org/10.2196/67891

32. McBain, R. K., Cantor, J. H., Zhang, L. A., Baker, O., Zhang, F., Burnett, A., Kofner, A., Breslau, J., Stein, B. D., Mehrotra, A., & Yu, H. (2025b). Evaluation of alignment between large language models and expert clinicians in suicide risk assessment. *Psychiatric Services*, *76*(11), 944–950. https://doi.org/10.1176/appi.ps.20250086

33. McBain, R. K., Bozick, R., Diliberti, M., Zhang, L. A., Zhang, F., Burnett, A., Kofner, A., Rader, B., Breslau, J., Stein, B. D., Mehrotra, A., Pines, L. U., Cantor, J., & Yu, H. (2025c). Use of generative AI for mental health advice among US adolescents and young adults. *JAMA Network Open*, *8*(11), e2542281. https://doi.org/10.1001/jamanetworkopen.2025.42281

34. McCoy, T. H., & Perlis, R. H. (2025). Reasoning language models for more transparent prediction of suicide risk. *BMJ Mental Health*, *28*(1), e301654. https://doi.org/10.1136/bmjment-2025-301654

35. Pichowicz, W., Kotas, M., & Piotrowski, P. (2025). Performance of mental health chatbot agents in detecting and managing suicidal ideation. *Scientific Reports*, *15*(1), 31652. https://doi.org/10.1038/s41598-025-17242-4

36. Prochaska, J. J., Vogel, E. A., Chieng, A., Kendra, M., Baiocchi, M., Pajarito, S., & Robinson, A. (2021). A therapeutic relational agent for reducing problematic substance use (Woebot): Development and usability study. *Journal of Medical Internet Research*, *23*(3), e24850. https://doi.org/10.2196/24850

37. Ruger-Navarrete, A., Gómez-Ferrera, M., Mérida-Yáñez, B., Vázquez-Lara, J. M., Gómez-Salgado, J., García-Oliva, S., Vázquez-Lara, M. D., Rodríguez-Diaz, L., Antúnez-Calvente, I., & Fernández-Carrasco, F. J. (2026). Artificial intelligence in the prevention and early detection of postpartum depression: A systematic review and meta-analysis. *Frontiers in Psychiatry*, *16*, 1734102. https://doi.org/10.3389/fpsyt.2025.1734102

38. Sohn, J.-S., Ha, B.-G., Park, S., Kim, J.-J., Lee, E., Oh, H., Lee, S., & Kim, E. (2026). Systematic review and meta analysis of chatbots in the management of depressive and anxiety symptoms. *npj Digital Medicine*, *9*(1), 377. https://doi.org/10.1038/s41746-026-02566-w

39. Swaminathan, A., López, I., Mar, R. A. G., Heist, T., McClintock, T., Caoili, K., Grace, M., Rubashkin, M., Boggs, M. N., Chen, J. H., Gevaert, O., Mou, D., & Nock, M. K. (2023). Natural language processing system for rapid detection and intervention of mental health crisis chat messages. *npj Digital Medicine*, *6*(1), 213. https://doi.org/10.1038/s41746-023-00951-3

40. Thomas, J., Lucht, A., Segler, J., Wundrack, R., Miché, M., Lieb, R., Kuchinke, L., & Meinlschmidt, G. (2025). An explainable artificial intelligence text classifier for suicidality prediction in youth crisis text line users: Development and validation study. *JMIR Public Health and Surveillance*, *11*, e63809. https://doi.org/10.2196/63809

41. Villarreal-Zegarra, D., Reategui-Rivera, C. M., García-Serna, J., Quispe-Callo, G., Lázaro-Cruz, G., Centeno-Terrazas, G., Galvez-Arevalo, R., Escobar-Agreda, S., Dominguez-Rodriguez, A., & Finkelstein, J. (2024). Self-administered interventions based on natural language processing models for reducing depressive and anxiety symptoms: Systematic review and meta-analysis. *JMIR Mental Health*, *11*, e59560. https://doi.org/10.2196/59560

42. Weber, S., Klebe, D., Wolf, L., Aeberli, C., Homan, S., Psathakis, N., Casanova, A., Macias, J. G., Sykstus, J., Li, J.-M., Provaznikova, B., Rohde, J., Frühschütz, L., Rühlmann, C., Welt, T., Kowatsch, T., Kleim, B., MULTICAST consortium, & Olbrich, S. (2026). *Suicide- and crisis-risk detection using large language models in mental-health chatbots* [Preprint]. medRxiv. https://doi.org/10.64898/2026.01.12.26343914

43. Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: A narrative review. *npj Digital Medicine*, *5*(1), 46. https://doi.org/10.1038/s41746-022-00589-7