International Journal of Engineering Research-Online
A Peer Reviewed International Journal
ISSN: 2321-7758 http://www.ijoer.in editorijoer@gmail.com
Vol.14., Issue.1, 2026
Jan-March.

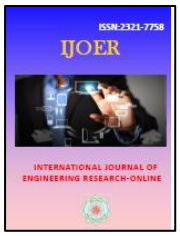
REVIEW ARTICLE

ISSN: 2321-7758

Conversational AI Chatbots for Mental Health Monitoring and Support: A Review of Architectures, Applications, Challenges, and Future Directions

Jayram A. Patel
Lecturer, Department of Electrical Engineering and IT, Polytechnic,
The M.S. University of Baroda, Vadodara, Gujarat, India

DOI: 10.33329/ijoer.14.1.15



Abstract

Conversational artificial intelligence (AI) has emerged as an important digital-health approach for delivering accessible mental-health information, psychoeducation, self-management guidance and structured interventions. Early conversational systems were predominantly rule-based and depended on pattern matching, scripts and dialogue trees. The field has subsequently progressed through machine learning, deep learning, transformer architectures and large language models (LLMs), enabling increasingly flexible and context-sensitive interaction. This review examines the evolution of mental-health conversational agents, their architectures, application areas, evaluation practices, and the safety, privacy and ethical issues that influence their deployment. Evidence from systematic reviews, meta-analyses and representative clinical studies indicates that conversational agents can provide short-term improvements in selected outcomes including depressive symptoms, anxiety, distress and well-being. Nevertheless, the evidence is heterogeneous, user attrition remains a practical limitation, and long-term clinical effectiveness is less established. Recent generative-AI systems further increase the importance of hallucination control, transparency, accountability and crisis-sensitive response mechanisms. The review identifies a research gap at the intersection of continuous mental-health monitoring, multimodal sensing, personalized conversation, risk-aware decision logic, privacy-preserving local processing and real-time Edge-AI deployment. A layered conceptual framework is proposed in which conversational inputs and wearable/mobile signals are subjected to local data quality checks, feature extraction, state estimation, longitudinal analysis, safety assessment and controlled response generation. The framework treats conversational AI as one component of a broader monitoring-and-support loop rather than as an independent replacement for clinical care. The review concludes that future systems should combine conversational intelligence with clinically grounded evaluation, explicit safety mechanisms, transparent architecture reporting, privacy-aware edge processing and human oversight.

Keywords: Conversational AI, Mental Health Chatbot, Large Language Model, Digital Mental Health, Mental Health Monitoring, Edge AI, Natural Language Processing, Privacy, Safety, Personalization.

1. INTRODUCTION

Mental health is increasingly recognized as an area in which digital technologies can supplement conventional services. Access to qualified professionals may be constrained by cost, geographic distribution, waiting time, stigma, workforce shortages and the practical difficulty of obtaining support at the moment a person experiences distress. Digital mental-health systems have therefore been investigated as mechanisms for providing information, self-management resources, monitoring, screening and structured interventions between or outside conventional clinical encounters. [1,2,3,7,10]

Conversational agents occupy a distinctive position within this digital-health landscape because the principal interaction takes place through natural-language dialogue. Instead of navigating menus or completing a static questionnaire, a user can describe a concern, answer a sequence of questions, receive an explanation or practice a structured technique through a dialogue interface. The conversational format can make digital interventions appear more accessible and can potentially support repeated engagement. However, a natural conversation should not be confused with clinical competence. The system must still be evaluated for validity, safety, reliability, privacy and its effect on meaningful health outcomes. [1,2,3,7,10]

The research community has approached mental-health chatbots from multiple directions. Computer-science studies commonly investigate natural-language processing, classification, dialogue management, language generation, model evaluation and user interaction. Clinical and medical studies more frequently investigate intervention design, symptom reduction, feasibility, therapeutic alliance, adherence and health outcomes. Cho et al. [1] described this interdisciplinary separation and emphasized the importance of connecting technical and medical perspectives. This observation is particularly relevant for emerging LLM-based systems, where improvements in language generation can occur much faster than clinical validation. [10]

Clinical evidence provides a qualified basis for optimism. He et al. [2] reviewed randomized controlled trials of conversational agents and reported small-to-moderate short-term effects for several mental-health outcomes. Li et al. [3] likewise found beneficial effects for depression and distress in AI-based conversational-agent interventions. Earlier work such as the Woebot randomized trial demonstrated the feasibility of a fully automated conversational CBT intervention for young adults with symptoms of depression and anxiety. These studies establish that conversational agents can be more than technical demonstrations, but they do not establish that every chatbot or every generative model is clinically effective.

At the same time, the technical architecture of conversational agents is changing rapidly. Rule-based systems emphasize control and predictability; machine-learning systems introduce data-driven classification and personalization; transformer models provide improved contextual representation; and LLMs can generate responses rather than selecting only from a prewritten library. This progression creates an engineering trade-off between flexibility and control. More generative systems can handle a broader range of language but can also produce unexpected, incorrect or unsafe responses. [1,2,3,7,10]

An additional challenge arises when mental-health support is combined with continuous sensing. Wearable and mobile devices can collect physiological or behavioral indicators that may provide context about activity, sleep-related patterns, heart-rate dynamics or other states. Such signals do not constitute a direct diagnosis of mental health, but they can provide additional information for longitudinal monitoring. Combining these signals with conversational information may enable a richer representation of a user's changing state, provided that the system explicitly manages uncertainty and avoids overinterpretation. [1,2,3,7,10]

Edge-AI is relevant to this problem because continuous monitoring can generate frequent and sensitive data. Local filtering, feature extraction and inference can reduce the amount of raw information that must be transmitted to remote services. Edge processing can also reduce latency and allow selected functions to operate during temporary loss of connectivity. These advantages are balanced by limited device resources, model-size constraints, energy consumption, update management and the need for robust local security. [1,2,3,7,10]

The objective of this review is to examine the development of mental-health conversational agents, compare their principal architectures, summarize their applications and evidence, identify recurring evaluation and ethical challenges, and derive a research gap relevant to a real-time Edge-AI-driven conversational framework for mental-health monitoring and support. [1,2,3,7,10]

Literature Review

1. Traditional Mental-Health Chatbots

Early conversational systems used pattern matching, predefined scripts and manually constructed dialogue trees. ELIZA is a well-known historical example: it simulated a psychotherapeutic conversation through linguistic patterns but did not possess clinical reasoning. Contemporary rule-based systems are more structured and can incorporate psychoeducation, cognitive-behavioral exercises, symptom questionnaires, reminders and referral instructions. [1,12]

The principal advantage of a rule-based architecture is predictability. A developer can define the permitted response space and explicitly prevent the system from making unsupported clinical claims. This is valuable in high-risk functions such as screening questions, crisis messages, referral information and structured therapeutic exercises. It also makes testing comparatively straightforward because each rule and transition can be inspected. [1,12]

The main limitation is rigidity. Human language is variable, and users frequently express the same concern using different vocabulary, grammar, cultural references and levels of detail. A dialogue tree can fail when the user's response does not match an expected branch. Repeated exposure to identical scripts may also reduce engagement. Consequently, rule-based methods remain useful as controlled components within larger systems rather than necessarily serving as the sole conversational architecture. [1,12]

2. Machine-Learning and Deep-Learning Chatbots

Machine-learning methods introduced data-driven approaches for intent classification, sentiment analysis, emotion recognition, entity extraction, response ranking and user-state estimation. Instead of requiring every linguistic expression to be explicitly programmed, the system learns patterns from training examples. Recurrent neural networks and long short-term memory models improved the ability to represent sequential information, while transformer-based encoders provided stronger contextual representations. [1,3,7,14]

In mental-health applications, machine-learning components may classify symptom-related language, identify emotional

states, estimate engagement, detect changes in interaction patterns or assist in risk screening. Such components can be placed before a dialogue manager so that the conversational response is conditioned on an estimated user state. Importantly, a model's output should be treated as an estimate with uncertainty rather than as an unquestionable fact. [1,3,7,14]

The quality of machine-learning systems depends heavily on training data. Mental-health language varies by age, culture, language, socioeconomic context and individual communication style. Dataset imbalance can therefore create systematic performance differences. Models can also experience distribution shift when deployed in settings that differ from their training environment. Validation across representative populations is consequently essential before clinical use. [1,3,7,14]

3. Transformer and Large Language Model-Based Systems

Transformer architectures changed conversational AI by allowing models to represent relationships among tokens across a conversation and to use broad pretraining corpora. LLMs can generate natural-language responses, summarize user statements, reformulate information and maintain conversational continuity. These properties are attractive for mental-health support because users may prefer a flexible dialogue over a rigid sequence of buttons and menus. [7,9,10,15,18]

The 2025 systematic review by Hua et al. [7] described the evolution of mental-health chatbots from rule-based systems through machine-learning systems to LLM-based systems. The review identified a rapid expansion of LLM research but also emphasized that many studies remained at technical validation or feasibility stages. This distinction is important: an LLM can demonstrate strong general conversational performance without demonstrating that its mental-health responses are safe, effective or appropriate for a vulnerable population.

Generative models introduce hallucination risk, in which fluent language may contain inaccurate or unsupported information. They may also respond with excessive confidence, misunderstand context, miss subtle indications of crisis or generate advice that is inappropriate for an individual. Prompting alone is therefore insufficient as a safety strategy. Evidence grounding, constrained response policies, explicit safety classifiers, escalation rules, uncertainty communication and continuous evaluation are more appropriate components of a health-oriented architecture. [7,9,10,15,18]

Another concern is anthropomorphism. Users may attribute understanding, empathy or authority to a system because it communicates fluently. In mental-health contexts this effect can be stronger because users may disclose sensitive information and seek emotional reassurance. System interfaces should therefore make the nature and limitations of the agent clear without making the interaction unnecessarily impersonal. [7,9,10,15,18]

4. Applications in Mental-Health Support

The literature describes a broad range of applications. Psychoeducation is one of the lower-risk uses, where a system provides structured information about symptoms, coping strategies and healthy routines. Self-management applications may guide users through behavioral activation, cognitive restructuring, breathing exercises, journaling or other structured activities. Screening-oriented systems can ask standardized questions and direct users toward appropriate services. [12,13,14,20]

Emotional-support applications attempt to provide empathic dialogue and a sense of availability. Their value may be particularly relevant when a user wants to express thoughts at a convenient time.

However, emotional support must be distinguished from psychotherapy and emergency care. The system should not imply that automated conversation can replace professional assessment or crisis intervention. [12,13,14,20]

Monitoring is another important application. A conversational agent can ask periodic questions about mood, stress, sleep, motivation or functioning and compare responses over time. If combined with wearable or mobile data, it may also construct a longitudinal context. Such monitoring can potentially identify deviations from an individual's baseline, but thresholds must be validated and the system should avoid converting sensor changes directly into diagnostic conclusions. [12,13,14,20]

Navigation and referral provide another useful function. A chatbot can help a user understand when professional support may be appropriate and can present predefined pathways to qualified services. This human-in-the-loop role is safer than assigning the system unrestricted authority over high-risk decisions. In an integrated architecture, the conversational layer can therefore act as an interface to a broader decision-support process rather than functioning as the entire system. [12,13,14,20]

5. Representative Clinical Evidence

Fitzpatrick et al. [4] evaluated Woebot in a randomized controlled trial involving young adults with symptoms of depression and anxiety. The study demonstrated the feasibility of delivering a structured CBT-oriented intervention through a fully automated conversational agent and reported a significant reduction in depression symptoms for the intervention group relative to the information-control condition over the short study period.

He et al. [2] subsequently synthesized randomized controlled trials and found small-to-moderate short-term effects across depression, anxiety, distress, stress and quality-of-life or well-being measures. Their findings support the proposition that conversational agents can have measurable effects, while also showing that outcomes differ by target condition and intervention design.

Li et al. [3] conducted a systematic review and meta-analysis of AI-based conversational agents and found significant effects for depression and distress, although overall psychological well-being was not consistently improved. These findings suggest that the effect of a conversational intervention is outcome-specific rather than universally beneficial.

More recent evidence concerning generative-AI mental-health chatbots is promising but less mature. Zhang et al. [9] synthesized studies of generative-AI chatbots and reported a pooled positive effect while noting heterogeneity and methodological limitations. The implication is not that LLM-based systems have already been clinically established, but that they justify more rigorous trials and standardized safety evaluation. [9,19]

6. Evaluation of Clinical and User Outcomes

Evaluation is one of the most important weaknesses identified across the literature. Jabir et al. [5] found substantial variation in outcome measurement, with 32 studies using 203 measurement instruments across clinical, user-experience and technical outcomes. Many instruments appeared in only one study or had limited reporting of validity. Such variation makes it difficult to compare apparently similar interventions.

A robust evaluation framework should distinguish at least three layers. Technical validation should measure response accuracy, latency, robustness, classification performance, failure rates and safety-trigger sensitivity. Feasibility and user evaluation should examine usability, engagement, trust, therapeutic alliance where relevant, satisfaction, adherence and attrition. Clinical evaluation should use

validated outcome measures and appropriately designed controlled studies. [16]

Attrition is particularly important. A technically accurate chatbot that users abandon after a few interactions cannot deliver a sustained intervention. The systematic review of conversational-agent attrition reported meaningful dropout across trials and identified study duration and intervention characteristics as relevant factors. This supports the inclusion of engagement monitoring and low-burden interaction design in future systems. [6]

Longitudinal evaluation is also required. Short-term improvements may not persist, and users can change their behavior as the novelty of a conversational system decreases. A monitoring-oriented system should therefore evaluate not only pre/post symptom changes but also engagement trajectories, repeated measurements, false alerts, missed alerts, intervention burden and changes in user trust. [5,6,7,19]

7. Safety, Privacy and Ethical Issues

Mental-health conversations can contain information about emotions, relationships, trauma, medication, behavior and other highly sensitive matters. Privacy risk can arise at every stage: collection, transmission, storage, processing, model inference, analytics, third-party integration and deletion. A responsible system should apply data minimization and should collect only information that is necessary for the declared purpose. [8,11,17,18]

Security is equally important. Authentication, authorization, encryption, secure storage, controlled software updates and protection against unauthorized access should be considered from the beginning of development. In a wearable or edge system, the physical device itself becomes part of the security boundary. Loss or theft of a device should not expose raw mental-health information or unrestricted model artifacts. [8,11,17,18]

The 2025 scoping review by Rahsepar Meadi et al. [8] identified recurring ethical themes including privacy and confidentiality, safety and potential harm, responsibility and accountability, justice, explainability and transparency, empathy and anthropomorphization. These themes show that technical performance alone cannot establish responsible deployment.

The World Health Organization's guidance [11] on AI for health emphasizes human autonomy, well-being and safety, transparency and explainability, responsibility and accountability, inclusiveness and equity, and responsiveness and sustainability. These principles can be translated into engineering requirements. Users should understand what the system does, what it does not do, how data are handled and what happens when the system detects a potentially serious situation.

Crisis handling deserves explicit treatment. A general-purpose chatbot should not be expected to infer every crisis state correctly from free text. A safer architecture combines risk-sensitive language analysis with predefined escalation policies. When a high-risk condition is detected, the system should move from ordinary supportive dialogue toward an appropriate human or emergency pathway rather than continuing an unrestricted conversation. [8,11,17,18]

Transparency is also important when generative models are used. The interface should identify the agent as an AI system, avoid implying human consciousness or professional licensure, and communicate limitations. Explainability does not necessarily require exposing internal model parameters; it can instead include understandable reasons for alerts, data-use explanations, model/version records and auditable decision pathways. [8,11,17,18]

8. Research Gap Identified

The reviewed literature reveals a gap between conversational intelligence and

continuous, privacy-aware, real-time mental-health monitoring. Existing systems often concentrate on one principal objective: conversational intervention, usability, symptom screening, outcome measurement or technical language-model performance. These contributions are valuable but often remain separate. [7,10,15,18]

A more integrated system needs to combine conversational information with physiological or behavioral context, longitudinal analysis and explicit safety logic. The architecture must also account for privacy and connectivity constraints. A cloud-only design can provide substantial computational resources but may require sensitive data to leave the user environment and may introduce latency or dependence on network availability. [7,10,15,18]

The research opportunity is therefore not simply to construct a more fluent chatbot. It is to develop a dependable closed-loop system capable of acquiring multimodal information, performing local preprocessing and inference, estimating changes relative to an individual's history, interacting naturally, recognizing risk and selecting a controlled response. The system should remain a monitoring-and-support platform rather than presenting unvalidated estimates as medical diagnoses. [7,10,15,18]

A second gap concerns-controlled adaptation. Personalization can improve relevance, but unrestricted self-learning is inappropriate for safety-critical mental-health decisions. A future system should therefore separate personalization from clinical decision logic and should use versioning, validation, rollback and performance monitoring for model updates. [7,10,15,18]

Comparative Analysis of Major Chatbot Architectures

A comparative analysis of major chatbot architectures reveals distinct trade-offs in core mechanism, strengths, limitations, typical uses, and mental-health suitability [5,6,7,18,19]. Rule-based systems rely on scripts, rules, and dialogue trees, offering predictability, auditability, and controllability, but they are rigid and limited in language coverage; they are typically used for screening, psychoeducation, and structured CBT, making them highly suitable for bounded tasks [5,6,7,18,19]. ML and deep learning approaches use intent classification and learned representations, providing adaptive and data-driven behaviour, though they suffer from dataset bias and validation burden; they are commonly applied to emotion and state classification and personalization, with moderate-to-high mental-health suitability provided adequate validation [5,6,7,18,19]. Transformer architectures employ contextual encoding and ranking, enabling better context handling, but they demand significant computational and data resources; they are used for classification, retrieval, and dialogue management, and are highly suitable for defined functions [5,6,7,18,19]. LLM and generative AI systems generate contextual text, offering flexible and natural dialogue, yet they raise concerns around hallucination, safety, and opacity; they are typically used for support, education, and coaching, and are promising for mental health when appropriate safeguards are in place [5,6,7,18,19]. Finally, hybrid edge-AI architectures combine local sensing with ML, LLMs, and rules, delivering privacy, low latency, and safety layering, at the cost of higher engineering complexity; they are well suited to real-time monitoring and support and represent a strong research direction for mental-health applications [5,6,7,18,19].

9. Proposed Conceptual Framework for Edge-AI Conversational Mental-Health Support

The proposed framework is based on a layered architecture in which conversational intelligence is integrated with sensing, local analytics and safety-aware decision logic. Two principal input streams are considered. The first consists of direct user interaction such as text, voice-derived information, questionnaire

responses and conversational history. The second consists of physiological or behavioral signals acquired from wearable or mobile sensors. The system does not assume that either stream alone is sufficient; instead, it combines them as complementary context. [1,7,10,11,17,18]

The first processing layer performs data quality assessment. Sensor data can contain missing samples, artifacts, motion effects, signal dropouts or battery-related interruptions. Conversational inputs can also be incomplete or ambiguous. Quality indicators should therefore accompany downstream features so that the decision layer knows when confidence should be reduced. [1,7,10,11,17,18]

The second layer performs feature extraction and local state estimation. Depending on the available sensors, features may include time-domain or frequency-domain physiological characteristics, activity summaries, interaction frequency, linguistic features and user-reported states. The output is a set of indicators rather than a direct diagnosis. [1,7,10,11,17,18]

The third layer maintains longitudinal context. A single observation may be less informative than a sustained deviation from an individual's baseline. Trend analysis can compare current estimates with historical ranges and can identify persistent changes, sudden anomalies or repeated patterns. Such analysis must account for missing data and normal variation. [1,7,10,11,17,18]

The fourth layer is the conversational engine. The engine receives relevant context and selects a response strategy. Responses may include a brief check-in, psychoeducation, a self-management activity, a request for additional information, or a recommendation to seek professional support. Generative language can be used within a controlled response policy. [1,7,10,11,17,18]

The fifth layer is safety and escalation. It operates independently enough to prevent a generative response from bypassing critical safeguards. Risk-sensitive inputs can trigger a predefined pathway, require confirmation, or route the user toward human support. The system should log the decision context in a privacy-aware manner so that safety events can be reviewed. [1,7,10,11,17,18]

The sixth layer is controlled adaptation. Personalization can use local history to adjust interaction timing, preferred language, response length or selected self-management content. Any model update that changes inference or safety behavior should be versioned and validated before deployment. This separates user-specific adaptation from uncontrolled alteration of the system's safety rules. [1,7,10,11,17,18]

INPUTS [1,7,10,11,17,18]

EDGE-AI PROCESSING [1,7,10,11,17,18]

OUTPUTS [1,7,10,11,17,18]

Wearable signals
User text / voice
Questionnaires [1,7,10,11,17,18]

Data quality checks
Filtering / preprocessing
Feature extraction [1,7,10,11,17,18]

Cleaned data
Quality indicators
Context features [1,7,10,11,17,18]

Current state + history
Personal profile [1,7,10,11,17,18]

Local state estimation
Longitudinal trend analysis [1,7,10,11,17,18]

State indicators
Trend scores
Anomaly indicators [1,7,10,11,17,18]

Safety rules
Interaction context [1,7,10,11,17,18]

Risk assessment
Decision logic
Controlled adaptation [1,7,10,11,17,18]

Response strategy
Escalation state [1,7,10,11,17,18]

Conversational context [1,7,10,11,17,18]

Response generation
Safety guardrails
Uncertainty handling [1,7,10,11,17,18]

Personalized support
Referral / human support [1,7,10,11,17,18]

Local storage / sync [1,7,10,11,17,18]

Versioning
Audit trail
Secure synchronization [1,7,10,11,17,18]

Authorized summaries
Alerts when required [1,7,10,11,17,18]

Figure 1. Proposed layered architecture for real-time Edge-AI conversational mental-health monitoring and support. [1,7,10,11,17,18]

10. Edge-AI Processing and Privacy-Preserving Deployment

Edge-AI refers to performing selected computation close to the source of data rather than sending every raw observation to a remote server. For a wearable mental-health monitoring system, local processing can include signal filtering, feature extraction, anomaly detection, model inference, trend computation and selected decision logic. The approach is especially attractive when the system operates continuously. [11,17,18]

Privacy is one of the strongest motivations for local processing. Raw sensor streams and conversational content can reveal more information than is necessary for a particular decision. If the device can transform raw data into a small set of features or risk indicators locally, the amount of sensitive information transmitted can be reduced. This does not eliminate privacy risk, because local storage and synchronization still require protection, but it can reduce unnecessary exposure. [11,17,18]

Latency is a second advantage. A local model does not need to wait for a round trip to a cloud service for every inference. This can be useful for immediate feedback and for safety-related functions where delayed processing is undesirable. Local operation also supports resilience when the network is unavailable, although remote communication may still be appropriate for authorized synchronization or escalation. [11,17,18]

Resource constraints are the principal engineering challenge. Wearable and embedded devices typically have limited memory, processing capability, storage and battery capacity compared with desktop or cloud systems. Models therefore need appropriate compression, quantization, efficient feature representations and carefully selected inference schedules. The architecture should also distinguish continuous low-cost processing from occasional high-cost operations. [11,17,18]

Security must be treated as part of the edge architecture. Local storage should be protected, software updates should be authenticated, access to diagnostic interfaces should be restricted and sensitive logs should be minimized. Model files can also require protection because unauthorized modification of a safety-critical model could alter system behavior. [11,17,18]

Offline operation introduces another design question: what happens when synchronization fails? The device should continue safe local operation where possible, queue only authorized information, prevent unbounded storage growth and synchronize when a trusted connection becomes available. The recovery process should preserve timestamps and model versions so that later analysis can distinguish delayed data from real-time data. [11,17,18]

11. Multimodal and Longitudinal Mental-Health Monitoring

Conversational information and wearable signals provide different forms of evidence. A user statement can provide direct self-reported information about mood or stress, whereas physiological and behavioral signals provide indirect measurements. Neither should automatically dominate the other. A robust system should preserve the distinction between self-report, sensor-derived features and model-generated estimates. [3,7,15,20]

Multimodal fusion can occur at different stages. Early fusion combines features before prediction, while late fusion combines separate model outputs. A hybrid approach may be preferable when modalities have different sampling rates or quality characteristics. For example, conversational events may occur intermittently while wearable measurements may be collected continuously. A temporal context layer can reconcile these different time scales. [3,7,15,20]

Longitudinal monitoring is particularly relevant because mental state is dynamic. Baseline behavior differs across individuals, and a value that is unusual for one person may be normal for another. Personal baselines can therefore improve the usefulness of anomaly detection, provided that the baseline is not itself contaminated by sustained deterioration. [3,7,15,20]

Trend analysis should use multiple observations rather than reacting to isolated fluctuations. The system may classify patterns such as stable, improving, gradually worsening, rapidly changing or uncertain. These categories should be treated as monitoring indicators and not diagnostic labels. A safety policy can then define what type of interaction is appropriate for each state. [3,7,15,20]

12. Personalization and Controlled Adaptation

Personalization can occur at the interaction level without changing the underlying clinical safety policy. Examples include preferred language, communication style, reminder timing, response length, selected educational resources and the amount of context presented to the user. Such personalization may improve engagement because users differ substantially in how they prefer to interact with digital systems. [1,7,10,16]

Controlled local adaptation can also use user-collected data to refine non-critical model components. However, adaptation should be bounded. A system should not automatically change a crisis threshold or rewrite safety rules simply because a local dataset produces a different statistical pattern. High-risk functions should remain under explicit governance. [1,7,10,16]

Version control is therefore essential. Every model or decision component should have a version identifier, validation status and rollback mechanism. Before deployment of an update, performance can be compared with the previous version using held-out data or predefined test scenarios. Post-update monitoring can detect degradation and initiate rollback. [1,7,10,16]

Personalization must also respect privacy. Local adaptation provides an opportunity to keep user-specific data on the device rather than uploading it for centralized training. Where external synchronization is required, only authorized information should be transferred and the purpose of the transfer should be explicit. [1,7,10,16]

13. Safety-Aware Conversational Decision Logic

A safety-aware chatbot should separate response generation from response authorization. The language model can

propose a response, but a safety layer should determine whether the response is permitted in the current context. This design reduces the possibility that a fluent generative model bypasses a critical constraint. [8,11,18]

Decision logic can use multiple signals: explicit user statements, risk-related language, repeated deterioration indicators, questionnaire results, abrupt behavioral changes and system confidence. The logic should also consider data quality. A low-quality sensor signal should not be interpreted as a reliable indication of worsening mental health. [8,11,18]

Responses can be categorized into levels such as routine support, additional assessment, professional-support recommendation and urgent escalation. The exact thresholds require clinical validation and should not be invented solely from engineering assumptions. During research development, these categories can be implemented as configurable policies and evaluated in simulation before human-subject deployment. [8,11,18]

Safety testing should include adversarial and unexpected conversations, ambiguous statements, contradictory user information, prompt injection attempts, missing sensor data and communication failures. The goal is not to prove that the system can handle every possible situation, but to identify known failure modes and ensure that the system fails safely. [8,11,18]

Evaluation Requirements for a Future Integrated System

Evaluation of a future integrated system should be structured across multiple layers, each with its own key measures, guiding questions, and expected evidence [5,6,7,18,19]. At the technical layer, evaluation should focus on accuracy, latency, robustness, energy consumption, and memory use, asking whether local inference works reliably under realistic device constraints and drawing on benchmarks, stress tests, and failure analysis as evidence [5,6,7,18,19]. At the conversational layer, the emphasis shifts to relevance, coherence, safety, and response quality, with the central question being whether the agent responds appropriately to different contexts, evaluated through expert review, structured test sets, and safety scenarios [5,6,7,18,19]. The usability layer should assess usability, engagement, trust, and attrition, asking whether users will continue using the system without excessive burden, and relying on pilot studies, logs, and validated questionnaires [5,6,7,18,19]. At the clinical layer, evaluation should centre on validated outcomes, effect size, and adverse events, asking whether the intervention improves meaningful mental-health outcomes, with controlled clinical studies providing the necessary evidence [5,6,7,18,19]. Finally, at the deployment layer, privacy, security, reliability, and update control become paramount, framed by the question of whether the system can operate safely over long periods, and evidenced through security testing, audit records, and field evaluation [5,6,7,18,19].

14. Results and Observations

- Observation 1 — Architecture is evolving faster than clinical validation. The literature shows a clear transition from deterministic scripts toward machine-learning, transformer and LLM-based systems. The technical capability of these systems has increased rapidly, but many LLM studies remain at technical or feasibility stages. The evidence therefore supports a careful distinction between conversational capability and clinical efficacy. [2,3,5,6,7,9,18,19]
- Observation 2 — Conversational agents can improve selected short-term outcomes. Meta-analytic evidence supports beneficial effects for depression, distress, anxiety and related measures in selected populations. The magnitude and consistency of benefit vary, so the appropriate conclusion is that conversational agents are promising

digital interventions rather than universally effective treatments. [2,3,5,6,7,9,18,19]

- Observation 3 — Outcome measurement remains inconsistent. Different studies use different instruments, intervention durations and outcome definitions. This limits direct comparison and reinforces the need for validated, standardized measures covering technical performance, user experience and clinical outcomes. [2,3,5,6,7,9,18,19]
- Observation 4 — Engagement is a system-level requirement. Attrition demonstrates that technical availability does not guarantee sustained participation. Future systems should measure interaction frequency, completion rates, session duration, dropout, user burden and changes in engagement over time. [2,3,5,6,7,9,18,19]
- Observation 5 — Safety and privacy must be architectural requirements. Mental-health data are sensitive and generative systems can produce unsafe outputs. Safety classifiers, guardrails, escalation rules, data minimization, transparency, access control and accountability should therefore be incorporated from the beginning. [2,3,5,6,7,9,18,19]
- Observation 6 — Edge-AI is a promising integration point. Local processing can reduce unnecessary transmission of raw data, reduce latency and support operation under limited connectivity. Its main challenges are computational resources, energy, model management and security. [2,3,5,6,7,9,18,19]
- Observation 7 — Monitoring should be separated from diagnosis. Sensor-derived indicators and model scores can support monitoring and decision support, but they should not be represented as clinical diagnoses unless validated through appropriate clinical research. This distinction should be visible in both the user interface and system documentation. [2,3,5,6,7,9,18,19]
- Observation 8 — Human oversight remains important for high-risk situations. The safest role for a conversational agent is to support, monitor and guide rather than assume unrestricted clinical authority. Escalation pathways should connect the automated system with qualified human or emergency resources where appropriate. [2,3,5,6,7,9,18,19]

15. Hua Y, Na H, Li Z, Liu F, Fang X, Clifton D, Torous J. A scoping review of large language models for generative tasks in mental health care. npj Digital Medicine. 2025;8:230. doi:10.1038/s41746-025-01611-4.

The central finding of this review is that the next stage of mental-health conversational AI should be defined by integration and validation rather than by conversational fluency alone. A system that can generate natural language but cannot reliably estimate uncertainty, recognize risk, protect sensitive data or provide a controlled escalation pathway is unsuitable for high-stakes deployment. [7,10,11,15,18]

A hybrid architecture is therefore preferable for research and engineering. Deterministic components can establish safety-critical boundaries; machine-learning models can estimate states from structured and sensor-derived data; and generative models can provide flexible natural-language interaction within those boundaries. This separation of roles can improve auditability while preserving conversational adaptability. [7,10,11,15,18]

The proposed Edge-AI orientation addresses a practical limitation of fully cloud-dependent systems. Continuous monitoring may generate frequent data, some of which are sensitive and unnecessary to transmit in raw form. Local filtering, feature extraction and inference can reduce exposure and communication overhead. At the same time, synchronization, data retention, model updates and offline recovery must be explicitly engineered. [7,10,11,15,18]

Clinical translation remains the most important unresolved issue. Technical evaluation should be followed by usability and feasibility studies and then appropriately designed clinical-efficacy studies. Hua et al. [7] proposed a three-tier evaluation logic consisting of technical validation, pilot feasibility/user engagement and clinical efficacy. This staged approach is particularly appropriate for an integrated Edge-AI system because it reduces the risk of moving an unvalidated architecture directly into high-stakes use.

The framework should also distinguish monitoring from diagnosis. A computational estimate of a user's state should not automatically be presented as a medical diagnosis. Outputs should be communicated as monitoring indicators, supportive feedback or escalation recommendations unless a clinically validated diagnostic workflow has been established. [7,10,11,15,18]

Another important implication is that future research should report architecture details more transparently. Publications should specify where data are processed, what model types are used, whether a response is retrieved or generated, what safety mechanisms exist, how failures are handled, what outcome measures are used and how user data are protected. Transparent reporting would improve reproducibility and allow researchers to compare systems more meaningfully. [7,10,11,15,18]

The practical value of the proposed framework lies in treating the chatbot as one component of a larger cyber-physical digital-health system. The wearable or mobile device provides context; the Edge-AI layer performs local computation; the conversation engine provides an accessible interface; the safety layer constrains behavior; and human support remains available when automated support is insufficient. This division creates a pathway from technical sensing to actionable support without assuming that any single algorithm can solve the complete mental-health problem. [7,10,11,15,18]

Comparative Analysis of Major Chatbot Architectures

A comparative analysis of major chatbot architectures reveals distinct trade-offs in core mechanism, strengths, limitations, typical uses, and mental-health suitability. Rule-based systems rely on scripts, rules, and dialogue trees, offering predictability, auditability, and controllability, but they are rigid and limited in language coverage; they are typically used for screening, psychoeducation, and structured CBT, making them highly suitable for bounded tasks. ML and deep learning approaches use intent classification and learned representations, providing adaptive and data-driven behaviour, though they suffer from dataset bias and validation burden; they are commonly applied to emotion and state classification and personalization, with moderate-to-high mental-health suitability provided adequate validation. Transformer architectures employ contextual encoding and ranking, enabling better context handling, but they demand significant computational and data resources; they are used for classification, retrieval, and dialogue management, and are highly suitable for defined functions. LLM and generative AI systems generate contextual text, offering flexible and natural dialogue, yet they raise concerns around hallucination, safety, and opacity; they are typically used for support, education, and coaching, and are promising for mental health when appropriate safeguards are in place. Finally, hybrid edge-AI architectures combine local sensing with ML, LLMs, and rules, delivering privacy, low latency, and safety layering, at the cost of higher engineering complexity; they are well suited to real-time monitoring and support and represent a strong research direction for mental-health applications (Beatty et al., 2022).

The evidence base is changing rapidly. Generative AI models, evaluation methods, deployment practices and regulatory expectations can change faster than conventional publication cycles. The literature cut-off of 28 February 2026 was therefore used to establish a defined scope for this manuscript.

Future research should first establish a reproducible search and evaluation protocol. The next stage should then develop and benchmark individual modules: sensor preprocessing, feature extraction, state estimation, longitudinal trend analysis, conversation management and safety classification. Module-level validation can identify failure sources before the complete system is tested.

Subsequent research should examine multimodal fusion of conversational and wearable data, privacy-preserving edge inference, validated mental-state estimation, longitudinal anomaly analysis, safety-aware response generation, controlled local adaptation, human-in-the-loop escalation and multilingual interaction. Special attention should be given to populations whose language, culture or digital access differs from the training environment.

Clinical evaluation should eventually use prospective study designs with validated mental-health outcome measures and clearly defined safety endpoints. The system should be evaluated not only for improvement but also for unintended consequences, false reassurance, inappropriate escalation, user over-reliance and privacy incidents.

Deployment research should also evaluate energy consumption, offline resilience, update reliability, device failure, model drift and secure synchronization. These engineering properties can determine whether an apparently successful laboratory prototype remains reliable during long-term real-world use.

Conversational AI has progressed from scripted mental-health chatbots to increasingly adaptive machine-learning, transformer and LLM-based systems. The reviewed evidence indicates that conversational agents can provide meaningful short-term support for selected mental-health outcomes, but the field continues to face heterogeneous evaluation, user attrition, limited long-term evidence, privacy risk and insufficient clinical validation of newer generative systems (May & Denecke, 2022).

The principal research opportunity is to move beyond isolated chatbot functionality toward an integrated, safety-aware and privacy-conscious monitoring-and-support platform. A hybrid Edge-AI architecture combining local sensing, data quality assessment, feature extraction, mental-state estimation, longitudinal analysis, conversational interaction, safety assessment and controlled adaptation provides a technically coherent direction for this next stage. [2,3,7,10,18,19]

Such a system should not be positioned as an autonomous replacement for professional mental-health care. Instead, it can serve as a low-threshold digital support layer that observes relevant changes, communicates with the user, provides evidence-based assistance and initiates appropriate escalation when predefined safety conditions are met. Its effectiveness must be demonstrated progressively, from technical validation to feasibility and finally clinical efficacy. [2,3,7,10,18,19]

The combination of conversational AI with Edge-AI is therefore best understood as an interdisciplinary engineering and clinical research problem. Success will depend not only on more capable language models, but also on reliable sensing, privacy-preserving computation, transparent decision logic, standardized evaluation, responsible adaptation and human oversight. These

requirements provide a foundation for future development of an Edge-AI-driven conversational framework for real-time mental-health monitoring and support. [2,3,7,10,18,19]

A clinically useful conversational mental-health system must satisfy requirements that are broader than natural-language generation. At the sensing layer, the system should tolerate missing samples, signal artifacts, temporary sensor disconnection and changes in device placement. At the inference layer, the model should provide stable outputs under resource constraints and should expose an indication of confidence or data quality. At the interaction layer, the system should maintain conversational context without retaining unnecessary sensitive information. These requirements become more important when the system is expected to operate continuously rather than only during scheduled chatbot sessions (Parks et al., 2025).

The software architecture should therefore be modular. A modular design allows the sensing, preprocessing, inference, dialogue, safety and storage components to be tested independently. It also makes it possible to replace one model without redesigning the entire system. For example, a lightweight local classifier may initially be used for state estimation, while a later version can introduce a more capable model after validation. Modular boundaries also simplify logging and failure analysis because the developer can determine which component produced a particular output. [11,17,18]

Resource management is a major consideration for Edge-AI. Continuous sensing can consume energy even when no conversational interaction is occurring. A practical system can use different processing frequencies for different tasks. Low-cost signal quality checks may run continuously, feature extraction may run at a moderate interval, and more computationally intensive analysis may be activated only when sufficient evidence of a meaningful change is present. Such event-driven processing can reduce energy consumption while preserving monitoring capability. [11,17,18]

Data storage should also be designed around purpose. Raw signals may be useful during model development but may not be necessary for every long-term deployment. A device can retain selected summaries, quality indicators and model outputs while applying a defined retention policy to raw data. During research, however, additional data may be retained under appropriate consent and security controls so that algorithms can be evaluated. The architecture should therefore distinguish research-mode storage from normal operational storage. [11,17,18]

Reliability in a mental-health monitoring system cannot be represented by a single accuracy value. A system may have high average classification accuracy while still failing in rare but important cases. Evaluation should therefore consider false positives, false negatives, missing data, delayed alerts, contradictory inputs and degraded operation. Safety-related functions should be tested separately because their acceptable failure characteristics differ from those of ordinary conversational responses (Feng et al., 2025).

A useful failure-mode analysis can begin with the input layer. If the wearable signal is unavailable, the system should recognize that absence rather than interpreting it as a meaningful physiological change. If the user gives an ambiguous statement, the conversation layer should ask for clarification when appropriate instead of assuming a particular mental state. If the model confidence is low, the decision layer should be able to select a conservative response. If network connectivity fails, the device should continue only those functions that have been validated for offline operation. [7,11,18]

The system should also protect against cascading errors. A noisy sensor signal should

not automatically become a high-risk mental-state estimate, and that estimate should not automatically trigger an alarming conversational message. Intermediate quality checks and confidence thresholds can interrupt such cascades. Similarly, a generative model should not directly control critical actions without passing through a safety policy. Layered validation is therefore a central engineering principle. [7,11,18]

Risk management should continue after deployment. Model performance can change when user behavior, language, devices or operating conditions change. Monitoring should therefore include model drift, alert frequency, unusual output patterns, user complaints and safety incidents. A controlled update process should allow the research team to compare versions and return to a previously validated model if a new version performs poorly. [7,11,18]

The effectiveness of a mental-health chatbot depends partly on whether people are willing and able to use it. Human-centered design should begin with the recognition that users may interact while stressed, tired, distracted or emotionally vulnerable. Interfaces should therefore minimize unnecessary steps, use understandable language and avoid presenting excessive information during moments of distress. The system should allow users to pause, decline questions and understand why information is being requested (Yang et al., 2025).

Language and culture are also important. Mental-health expressions differ across languages and communities, and direct translation of clinical terminology may not preserve meaning. A system intended for diverse users should be evaluated with representative language samples and should avoid assuming that one communication style is universal. Multilingual capability should therefore be treated as a validation problem rather than simply a translation feature. [8,16,18]

Trust should be built through transparency rather than simulated humanity. Users can be informed that they are communicating with an AI system, what types of information are being processed and what the system can and cannot do. The interface should also distinguish automated suggestions from professional advice. This approach can reduce inappropriate reliance while still allowing the conversation to remain supportive and accessible. [8,16,18]

Personalization should be user-controlled wherever practical. Users may prefer different reminder frequencies, interaction styles or levels of detail. Allowing these preferences can improve usability without changing the safety-critical behavior of the system. Personalization should not, however, pressure users into disclosure or create a false impression that the system knows the user more deeply than the available evidence supports. [8,16,18]

Research Roadmap for the Proposed Framework

The literature review suggests a staged research roadmap. Stage I should establish the data and sensing pipeline. This includes sensor acquisition, timestamp management, data-quality assessment, local storage and preprocessing. The primary objective at this stage is dependable data acquisition rather than mental-health prediction. The output should be a reproducible dataset and a validated preprocessing pipeline. [7,10,11,15,18]

Stage II should investigate feature extraction and state estimation. Candidate features can be evaluated using offline datasets and controlled experiments before they are incorporated into a live monitoring system. The important question is not only whether a feature correlates with a target outcome, but also whether it remains reliable across different users and operating conditions. [7,10,11,15,18]

Stage III should integrate conversational inputs with the estimated state. At this point, the system can test whether additional context improves the relevance of responses and whether conversational information can complement sensor-derived indicators. The evaluation should include cases in which the two modalities disagree, because disagreement is expected in real-world use. [7,10,11,15,18]

Stage IV should introduce the safety and escalation layer. Safety scenarios should be constructed in collaboration with appropriate domain experts. The objective should be to define conservative and auditable behavior rather than to maximize automated decision making. Human review can be used to assess difficult cases and refine the decision policy. [7,10,11,15,18]

Stage V should evaluate controlled personalization and adaptation. Non-critical interaction parameters can be adapted to individual preferences and usage patterns, while high-risk decision rules remain fixed or subject to explicit validation. Model versioning, rollback and auditability should be included before any adaptive component is deployed outside a controlled research setting. [7,10,11,15,18]

Stage VI should evaluate the complete system longitudinally. This stage should examine technical reliability, user engagement, privacy, safety events and clinically meaningful outcomes over a sufficiently long period. The final objective is not merely to demonstrate that the components work individually, but to establish that the complete monitoring-and-support loop provides benefit without introducing unacceptable risk. [7,10,11,15,18]

Expected Research Contributions

Based on the identified gap, the proposed research direction can contribute at several levels. At the architectural level, it can provide a reference design for combining wearable sensing, Edge-AI inference, conversational interaction and safety logic. At the algorithmic level, it can investigate lightweight models and multimodal feature fusion suitable for resource-constrained devices. At the system level, it can address local storage, longitudinal analysis, controlled adaptation and offline operation. [1,7,10,15]

A further contribution can be a structured evaluation framework that separates technical, conversational, usability, safety and clinical dimensions. Such separation is valuable because a system may perform well in one dimension and poorly in another. Reporting these dimensions independently can make future studies easier to compare and can reduce the tendency to use a single headline metric as evidence of overall system quality. [1,7,10,15]

The research can also contribute to privacy-aware deployment principles for continuous monitoring. Instead of assuming that all data should be transferred to a central service, the proposed architecture can investigate which processing tasks are feasible locally and which information, if any, needs authorized synchronization. This provides an engineering basis for reducing data exposure while maintaining useful functionality. [1,7,10,15]

Finally, the work can contribute a bridge between conversational AI research and wearable Edge-AI research. The two areas are often investigated separately: conversational studies focus on language interaction, while wearable studies focus on sensing and classification. An integrated framework provides an opportunity to investigate how these streams can complement each other within a single, safety-aware digital-health platform. [1,7,10,15]

Key Research Challenges and Open Problems

Despite rapid progress, several open problems must be addressed before integrated conversational Edge-AI systems can be considered dependable for long-term mental-

health support. The first is the problem of ground truth. Mental state is multidimensional and changes over time, while wearable signals are indirect indicators. A model therefore requires carefully defined labels, repeated measurements and appropriate clinical instruments. Treating a single questionnaire score as a complete representation of mental state can oversimplify the problem.

The second challenge is multimodal synchronization. Conversational events are irregular, while sensor streams may be sampled continuously. A useful system must preserve timestamps, distinguish current observations from historical information and account for missing intervals. Incorrect temporal alignment can create misleading associations and can cause the system to attribute a later conversational state to an earlier physiological event.

The third challenge is generalization. A model trained on a limited population may perform differently when deployed with users of different ages, languages, cultures, lifestyles or devices. Generalization should therefore be evaluated explicitly. Where performance varies, the system should expose uncertainty and avoid presenting a weak estimate as a confident conclusion.

The fourth challenge concerns model drift and personalization. Human behavior changes over weeks and months, and sensor characteristics can also change because of device placement, firmware or environmental conditions. A static model may gradually lose relevance, but uncontrolled adaptation creates its own risks. Future research should therefore investigate controlled adaptation strategies that can detect drift, request additional validation and roll back changes when required.

The fifth challenge is the evaluation of safety in generative conversation. Standard language-quality metrics cannot capture all clinically relevant failure modes. Evaluation should include scenario-based testing, expert review, adversarial prompts, ambiguous language, contradictory information and simulated escalation cases. The objective should be to measure not only whether a response is fluent, but whether it is appropriate, bounded and safe for the context.

The sixth challenge is determining the appropriate division between local and remote computation. Some functions may be well suited to an embedded device, while larger language models may require more computational resources. A future hybrid system could use lightweight local models for continuous monitoring and reserve remote computation for authorized, non-critical tasks. Such an architecture must ensure that loss of remote connectivity does not disable essential safety functions.

Finally, there is a governance challenge. A system that continuously monitors a person and produces mental-health-related indicators requires clear responsibility for data handling, model updates, alerts and user support. Technical documentation should therefore record model versions, data pathways, safety policies and update decisions. Governance should evolve together with the system rather than being added after the technical architecture is complete.

References

[1]. Cho YM, Rai S, Ungar L, Sedoc J, Guntuku S. An integrative survey on mental health conversational agents to bridge computer science and medical perspectives. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023. p. 11346-11369. doi:10.18653/v1/2023.emnlp-main.698.

[2]. He Y, Yang L, Li T, Qian C, Su Z, Zhang Q, et al. Conversational agent interventions for mental health problems: systematic review and meta-analysis of randomized controlled trials. J Med Internet Res. 2023;25:e43862. doi:10.2196/43862.

[3]. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC, et al. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med. 2023;6:236. doi:10.1038/s41746-023-00979-5.

[4]. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health. 2017;4(2):e19. doi:10.2196/mental.7785.

[5]. Jabir AI, Martinengo L, Lin X, Torous J, Subramaniam M, Tudor Car L. Evaluating conversational agents for mental health: scoping review of outcomes and outcome measurement instruments. J Med Internet Res. 2023;25:e44548. doi:10.2196/44548.

[6]. Jabir AI, Lin X, Martinengo L, Sharp G, Theng YL, Tudor Car L. Attrition in conversational agent-delivered mental health interventions: systematic review and meta-analysis. J Med Internet Res. 2024;26:e48168. doi:10.2196/48168.

[7]. 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. doi:10.1002/wps.21352.

[8]. Rahsepar Meadi M, et al. Exploring the ethical challenges of conversational AI in mental health care: scoping review. JMIR Ment Health. 2025;12:e60432. doi:10.2196/60432.

[9]. Zhang Q, Zhang R, Xiong Y, Sui Y, Tong C, Lin FH. Generative AI mental health chatbots as therapeutic tools: systematic review and meta-analysis of their role in reducing mental health issues. J Med Internet Res. 2025;27:e78238. doi:10.2196/78238.

[10]. Khosravi M, Izadi R. Mental health chatbots and their technical features: a systematic review of reviews and a thematic analysis. Cambridge Prisms Glob Ment Health. 2026;13:e28. doi:10.1017/gmh.2026.10144.

[11]. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021. ISBN: 978-92-4-002920-0.

[12]. Abd-Alrazaq AA, Alajlani M, Alalwan AA, Bewick BM, Gardner P, Househ M. An overview of the features of chatbots in mental health: a scoping review. Int J Med Inform. 2019;132:103978. doi:10.1016/j.ijmedinf.2019.103978.

[13]. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res. 2020;22(7):e16021. doi:10.2196/16021.

[14]. Lin X, et al. Scope, characteristics, behavior change techniques, and quality of conversational agents for mental health and well-being: systematic assessment of apps. J Med Internet Res. 2023;25:e45984. doi:10.2196/45984.

[15]. Hua Y, Na H, Li Z, Liu F, Fang X, Clifton D, et al. A scoping review of large language models for generative tasks in mental health care. NPJ Digit Med. 2025;8:230. doi:10.1038/s41746-025-01611-4.

[16]. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health. 2022;4:847991. doi:10.3389/fdgth.2022.847991.

[17]. May R, Denecke K. Security, privacy, and healthcare-related conversational agents: a scoping review. Inform Health Soc Care. 2022;47(2):194-210. doi:10.1080/17538157.2021.1983578.

[18]. Parks A, Travers E, Perera-Delcourt R, Major M, Economides M, Mullan P. Is this chatbot safe and evidence-based? A call for the critical evaluation of generative AI mental health chatbots. J Particip Med. 2025;17:e69534. doi:10.2196/69534.

[19]. Feng Y, Hang Y, Wu W, Song X, Xiao X, Dong F, et al. Effectiveness of AI-driven conversational agents in improving mental health among young people: systematic review and meta-analysis. J Med Internet Res. 2025;27:e69639. doi:10.2196/69639.

[20]. Yang Q, Cheung K, Zhang Y, Zhang Y, Qin J, Xie YJ. Conversational agents in physical and psychological symptom management: a systematic review of randomized controlled trials. Int J Nurs Stud. 2025;163:104991. doi:10.1016/j.ijnurstu.2024.104991.