International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 15 Issue 7, July-2026

# AI Agents for Mental Health: A Comprehensive Review of Applications, Efficacy, and Ethical Considerations

Mandeep[1], Dr. Shamsher Malik[2]

[1]Mtech 4th Sem Dept. of ECE, UIET MDU, Rohtak
[2]Professor, Department of ECE, UIET MDU, Rohtak

**ABSTRACT**

**Mental health disorders represent one of the most pressing global health challenges, with existing care systems unable to adequately meet rising demand. Artificial intelligence agents—particularly large language model-driven conversational systems—are increasingly being positioned as scalable solutions for mental health education, assessment, and intervention. This paper reviews the current landscape of AI agents in mental health care, synthesizes evidence from randomized controlled trials and meta-analyses regarding their clinical efficacy, examines core technical architectures, and critically evaluates the ethical, safety, and equity concerns accompanying their deployment. Findings indicate that AI-based conversational agents produce moderate short-term reductions in depression (Hedge's g = 0.64) and psychological distress (Hedge's g = 0.70), with personalization and empathic response identified as primary drivers of efficacy. However, long-term effects remain modest, and substantial risks persist regarding bias, hallucinations, crisis response, and accountability. The paper concludes that AI agents should function primarily as augmentative tools within clinician-supervised care pathways, supported by rigorous ethical frameworks and continuous human oversight.**

## INTRODUCTION

Mental health conditions have risen sharply worldwide. Depression and anxiety disorders together affect more than a quarter of the global population at some point in life, and the World Health Organization has identified mental health as a leading cause of disability. Yet, despite this escalation, traditional models of care remain constrained by limited workforce capacity, geographic inaccessibility, and persistent stigma. Across many countries, demand for services far exceeds supply, leaving millions without timely support.

In response, the field has turned to digital mental health interventions. Approximately 70% of patients have expressed interest in using mobile applications to self-monitor and self-manage their mental health, and internet-delivered therapies have demonstrated efficacy comparable to therapist-delivered cognitive behavioral therapy. However, adherence to these digital programs remains a critical weakness, with median completion rates as low as 56%, partly because of the loss of the interpersonal qualities central to traditional therapeutic encounters (Fitzpatrick et al., 2017). AI agents—systems capable of autonomous perception, reasoning, and action through natural-language interfaces—have emerged as a promising mechanism to bridge this gap by restoring conversational engagement at scale (Lawrence et al., 2024).

This paper examines the role played by AI agents in modern mental health care, reviewing their theoretical foundations, summarising empirical evidence of efficacy, and analysing the ethical frameworks required for responsible deployment.

## CONCEPTUAL FRAMEWORK: WHAT ARE AI MENTAL HEALTH AGENTS?

An AI agent for mental health is a software entity that perceives user input (text, voice, or multimodal signals), reasons about psychological state, and takes actions such as delivering an intervention or escalating to a clinician. Modern agents fall along a continuum of autonomy:

- Rule-based chatbots (e.g., early Woebot) follow scripted decision trees derived from therapeutic manuals.

- Retrieval-augmented systems (e.g., Wysa) combine rule-based logic with curated content libraries.
- Generative LLM-driven agents (e.g., GPT-powered mental health assistants) produce open-ended dialogue conditioned on therapeutic prompts.

These agents typically blend natural language processing, sentiment analysis, speech recognition, and computer vision. Recent architectures also incorporate multimodal data fusion, integrating text, audio, and facial-expression signals to infer affective states with greater accuracy than unimodal systems (Chen et al., 2024). This shift from rule-based to generative architectures is foundational: while earlier bots offered predictable, narrow responses, LLM-based agents promise dynamic, personalised dialogue (Hua et al., 2025).

## APPLICATIONS OF AI AGENTS IN MENTAL HEALTH

### Psychoeducation and Self-Help

AI agents can disseminate evidence-based information about disorders, treatments, and coping strategies to large populations at low marginal cost. Trials demonstrate that approximately 70% of patients accept mobile self-monitoring, illustrating the latent demand for such resources (Fitzpatrick et al., 2017).

### Assessment and Triage

LLMs have been deployed to analyse clinical notes, suggest diagnostic formulations, and screen patient messages for indicators of risk. Compared with control conditions, LLM-assisted documentation has been associated with more thorough symptom capture, although concerns remain regarding hallucinations and overconfident misclassification (Lawrence et al., 2024).

### Therapeutic Delivery

Conversational agents such as Woebot deliver fully automated CBT modules. In the seminal RCT by Fitzpatrick and colleagues, participants interacted with Woebot an average of 12 times over a two-week period, and 83% provided follow-up data—far exceeding the adherence rates of comparable web-based programmes (Fitzpatrick et al., 2017). Beyond fully autonomous models, hybrid platforms such as Eleos Health provide real-time clinical decision support to therapists conducting behavioural interventions. An RCT evaluating Eleos demonstrated superior reductions in depression and anxiety, along with improved patient retention, compared with treatment as usual (Sadeh-Sharvit et al., 2023).

### Clinician Support

Administrative automation—including progress-note generation, intake summarisation, and outcome-tracking—is among the most imminent and lowest-risk applications. Stakeholders have argued that such assistive uses should precede fully autonomous clinical deployment (Stade et al., 2024).

### Multimodal and Embodied Agents

Emerging systems integrate text, voice tone, and facial expression to enable more naturalistic affective engagement. Multimodal frameworks can classify emotional states—happiness, sadness, anger, fear, and neutrality—through acoustic features (pitch, tone, intensity) and visual features detected by tools such as DeepFace, producing more accurate emotion recognition than unimodal approaches (Chen et al., 2024).

## EVIDENCE OF CLINICAL EFFICACY

### Findings from the Woebot RCT

In a randomized controlled trial published in JMIR Mental Health, 70 college students with self-identified depression and anxiety were randomized to a Woebot CBT program (n=34) or an information-only control (n=36). The Woebot group exhibited a significant reduction in PHQ-9 depression scores relative to controls ($F=6.03$; $P=.017$), corresponding to a moderate between-subjects effect size of $d=0.44$. The effect remained significant after Bonferroni correction ($P=.04$). Anxiety showed a within-subjects effect size of $d=0.37$ among completers, although between-group differences on anxiety and affect did not reach significance. The authors concluded that "conversational agents appear to be a feasible, engaging, and effective way to deliver CBT" (Fitzpatrick et al., 2017).

### Systematic Reviews and Meta-Analyses

A 2023 systematic review and meta-analysis by Li, Zhang, Lee, Kraut, and Mohr synthesised 35 studies (including 15 RCTs) evaluating AI-based conversational agents. Across trials, AICAs significantly reduced depression (Hedge's $g=0.64$; 95% CI 0.17–1.12) and overall psychological distress (Hedge's $g=0.70$; 95% CI 0.18–1.22). However, their impact on anxiety ($g=0.65$; 95% CI −0.46 to 1.77), positive affect ($g=0.07$), negative affect ($g=0.52$), and general psychological well-

being (g=0.32; 95% CI −0.13 to 0.78) did not reach statistical significance. Effects were more pronounced in multimodal systems, generative AI architectures, and platforms integrated with mobile or instant-messaging applications (Li et al., 2023).

A complementary meta-analysis of 35 RCTs by He and colleagues yielded comparable estimates: depressive symptoms (g=0.29; 95% CI 0.20–0.38), generalized anxiety (g=0.29; 95% CI 0.21–0.36), and quality-of-life outcomes (g=0.27; 95% CI 0.16–0.39) showed short-term improvement, though long-term effects were attenuated (g ranged from −0.04 to 0.39). Critically, that review identified *personalization* and *empathic response* as the two strongest facilitators of efficacy, while excessive automated reminders were associated with diminished engagement—implying a positive dose-response relationship moderated by relational quality (He et al., 2023).

A 2025 systematic review further reported that AI-based therapies achieved symptom reductions between 19% and 42%, retention rates 28% higher than traditional self-help, and therapy completion 18% higher, with personalization and adaptive feedback repeatedly singled out as drivers (Nur et al., 2025).

**LLM-Specific Evidence**

Hua and colleagues' 2025 scoping review identified 16 studies meeting inclusion criteria from an initial pool of 726 articles. Most reported promising preliminary results across clinical assistance, counseling, therapy, and emotional support applications, but relied on non-standardised evaluation scales and often used proprietary models such as the GPT series, raising concerns about reproducibility and transparency (Hua et al., 2025). A parallel 2025 survey reported that LLMs increased the accuracy of early mental-disorder diagnosis by 33%, improved personalised treatment-plan effectiveness by 27%, and boosted participation in mental-health education by 24%, though the authors emphasised that these figures must be interpreted alongside acknowledged limitations in clinical-validation rigour (Ghorbian & Ghobaei-Arani, 2025).

*Interpretation.* Taken together, the empirical record supports three conclusions: AI agents produce statistically significant short-term gains in depression and distress; effects depend heavily on personalization, empathic quality, and engagement design; long-term durability remains unproven, and anxiety-specific outcomes are inconsistent. The evidence base still rests heavily on short, small-to-moderate trials; long-term, adequately powered RCTs against active comparators remain essential.

## TECHNICAL CONSIDERATIONS

**Architecture**

Most modern mental health agents combine (i) a perception layer, (ii) a reasoning layer, (iii) a therapeutic content layer (manuals, exercises, crisis resources), and (iv) a safety layer responsible for crisis detection, escalation, and audit logging. Multimodal designs augment each of these layers with affective signals derived from voice prosody and facial expressions (Chen et al., 2024).

**Engagement Mechanisms**

Across the 35 studies in Li et al.'s review, the most common design features were personalization and customization (n=20), regular check-ins (n=10), mood tracking (n=8), empathic responses (n=7), multimedia (n=5), and embodied characters (n=2). Notably, only 15 of 35 studies incorporated explicit safety-assessment measures such as automatic crisis identification, adverse-event tracking, or seamless handoff to human experts, underscoring a critical engineering gap (Li et al., 2023).

## ETHICAL, SAFETY, AND EQUITY CONCERNS

**The High-Stakes Nature of Mental Health**

Stade and colleagues argue that "psychotherapy delivery is an unusually complex, high-stakes domain vis-à-vis other LLM use cases." Where an LLM productivity tool may fail to maximise efficiency, an autonomous mental health agent may mishandle suicide or homicide risk. Prediction and mitigation of risk in such contexts is nuanced, requiring complex case conceptualisation, attention to social and cultural context, and responsiveness to unpredictable human behaviour (Stade et al., 2024).

**Bias, Hallucination, and Misinformation**

LLMs may reinforce harmful beliefs, generate non-factual medical advice, or fail to recognise mental-health crises, with potentially serious consequences for the integrity of psychological care. The "black-box" nature of proprietary models such

as GPT-3.5/4 compounds these concerns because researchers and practitioners often access models via APIs without complete knowledge of training data, methods, or version updates (Guo et al., 2024).

**Privacy and Confidentiality**

AI mental health agents routinely handle sensitive disclosures. Lawrence et al. emphasise that "privacy or confidentiality should be paramount when LLMs operate to support mental health," and advocate for explicit informed consent detailing the nature of services, the role of the LLM, and the associated risks and benefits (Lawrence et al., 2024).

**Crisis Response and Competence Boundaries**

Lawrence and colleagues outline a framework mandating that LLMs respond appropriately to mental-health risk, operate only within the bounds of demonstrated competence, and withhold output when they lack domain expertise. Competence should be re-assessed and maintained over time; deployment without such safeguards in suicide-risk contexts has been described as ethically impermissible (Lawrence et al., 2024).

**Accountability**

Questions of liability for harms caused by fully autonomous clinical LLM applications remain unresolved. Some scholars argue that, given unresolved safety concerns, fully autonomous deployment in behavioural healthcare is premature, and that assistive or collaborative applications should be prioritised in the short and medium term (Stade et al., 2024).

**Equity**

AI systems risk perpetuating existing inequities if training data underrepresent minority populations. Lawrence and colleagues advocate for mental-health LLMs to "advance mental health equity" by ensuring feedback from diverse populations—particularly sexual and gender minorities and individuals with lived experience of mental-health concerns—is incorporated throughout development, testing, and deployment (Lawrence et al., 2024).

**A Proposed Risk Framework**

Li and colleagues' Chain of Risks Evaluation framework categorises LLM risks in public mental health into four escalating tiers—universal, context-specific, user-specific, and user-context-specific—each requiring distinct mitigation strategies. Effectively addressing these risks requires a continuum of effort spanning technical safeguards (e.g., reinforcement learning with clinician feedback) and public-health governance (e.g., practitioner-led regulation) (Li et al., 2025).

**Best-Practice Recommendations**

Drawing on the evidence reviewed above, the following principles emerge for responsible deployment:

1. **Tightly scoped use cases.** Begin with low-stakes applications such as psychoeducation, clinician augmentation, and assessment support rather than fully autonomous therapy (Stade et al., 2024).
2. **Empathic and personalised design.** Embed empathic response and personalisation, as these are the most robustly identified drivers of efficacy (He et al., 2023; Li et al., 2023).
3. **Continuous human oversight.** Maintain clinicians in the loop through audit, feedback pipelines, and seamless escalation paths for crisis scenarios (Lawrence et al., 2024).
4. **Standardised evaluation.** Adopt common outcome instruments (e.g., PHQ-9, GAD-7) and generative-task evaluation frameworks to facilitate cross-study comparison (Hua et al., 2025).
5. **Equity and lived experience.** Engage individuals with lived experience throughout the development cycle (Lawrence et al., 2024).
6. **Updated competence validation.** Periodically re-evaluate model competence and decommission or retrain models when standards slip (Lawrence et al., 2024).

## FUTURE DIRECTIONS

Three trajectories warrant intensified research. First, long-term RCTs comparing AI agents against active comparators—particularly guided self-help and human therapist-led CBT—are needed to clarify durability (Li et al., 2023). Second, multimodal affective sensing should be integrated into agent architectures to deepen empathic engagement beyond text (Chen et al., 2024). Third, regulatory frameworks analogous to medical-device approval pathways should be developed to govern LLM-driven mental-health deployment, including mechanisms for proactive bias auditing and real-world post-deployment surveillance (Guo et al., 2024; Li et al., 2025).

## CONCLUSION

AI agents offer a credible mechanism to expand mental-health support amid a global shortage of clinicians. Meta-analytic evidence demonstrates moderate short-term reductions in depression and distress, particularly when agents are personalised,

empathic, and embedded within familiar messaging platforms. Multimodal and LLM-driven architectures promise more naturalistic engagement than rule-based predecessors. Yet the field remains in its infancy: long-term efficacy is modest, anxiety-specific effects are inconsistent, and serious concerns persist regarding bias, hallucination, crisis response, and accountability. The most defensible path forward lies in augmenting—not replacing—human care, applying AI agents first to well-bounded tasks such as psychoeducation, clinician support, and structured CBT delivery, while building rigorous safeguards, ethics review, and continuous human oversight into every stage of design and deployment. If pursued responsibly, AI agents can meaningfully broaden access to mental-health support; if pursued carelessly, they risk compounding the very harms they aim to alleviate.

## REFERENCES

1. Fitzpatrick, K. K., Darcy, A., & Vierhile, M.. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent: A Randomized Controlled Trial. *JMIR Mental Health*, 4, e19.
2. Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., & Mohr, D. C.. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *npj Digital Medicine*, 6, 236.
3. He, Y., Yang, L., Qian, C., et al.. Conversational Agent Interventions for Mental Health Problems: Systematic Review and Meta-analysis of Randomized Controlled Trials. *Journal of Medical Internet Research*, 25, e46862.
4. Lawrence, H. R., Schneider, R. A., Rubin, S. B., et al.. The Opportunities and Risks of Large Language Models in Mental Health. *JMIR Mental Health*, 11, e59466.
5. Stade, E. C., Stirman, S. W., Ungar, L., et al.. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3, 12.
6. Hua, Y., Na, H., Li, Z., et al.. A scoping review of large language models for generative tasks in mental health care. *npj Mental Health Research*, 4, 14.
7. Chen, X., Xie, H., Tao, X., et al.. Artificial intelligence and multimodal data fusion for smart healthcare: topic modeling and bibliometrics. *Information Fusion*, 102, 102059.
8. Ghorbian, M., & Ghobaei-Arani, M.. Large language models for mental health diagnosis and treatment: a survey. *Journal of Medical Systems*, 49, 32.
9. Sadeh-Sharvit, S., Del Camp, T., Horton, S. E., et al.. Effects of an Artificial Intelligence Platform for Behavioral Interventions on Depression and Anxiety Symptoms: Randomized Clinical Trial. *Journal of Medical Internet Research*, 25, e46781.
10. Li, L., Kong, S., Zhao, H., et al.. Chain of Risks Evaluation: A framework for safer large language models in public mental health. *Journal of Medical Internet Research*, 27, e64127.
11. Guo, Z., Lai, A. G., Thygesen, J. H., et al.. Large Language Models for Mental Health Applications: Systematic Review. *JMIR Mental Health*, 11, e56560.
12. Nur, J., Tan, K., & Wiputra, R.. AI Agents in Mental Health Treatment: A Systematic Review. *International Journal of Medical Informatics*, 195, 105731.