Effectiveness of AI-Assisted Mental Health Interventions in Online Counselling: A Study of Chatbot-Guided Support and Ethical Considerations

Annanya Jayaswal

Amity University, Noida

[1] *Date of Receiving: 29/12/2024;*
Date of Acceptance: 21/02/2025;
Date of Publication: 16/03/2025

ABSTRACT

The integration of Artificial Intelligence (AI) into mental health services has introduced new paradigms in online counselling, particularly through the use of chatbot-guided interventions. This paper explores the effectiveness of AI-assisted mental health interventions, focusing on chatbot-supported counselling, while critically analysing the associated ethical considerations. The study synthesizes evidence from recent randomized controlled trials (RCTs), systematic reviews, and pilot programs evaluating the impact of AI-driven conversational agents (CAs) on mental health outcomes such as depression, anxiety, and psychological distress.

Meta-analytic findings suggest moderate effectiveness of chatbot-guided interventions, especially in reducing depressive and anxiety symptoms in users with mild to moderate mental health conditions. Specific trials, including those using large language model (LLM)–based chatbots and peer-support augmentation with AI, demonstrate promise in improving user engagement and emotional support delivery. However, the lack of human empathy and limitations in handling crisis situations highlight the need for hybrid models that combine human oversight with chatbot support.

The paper also presents a comparative analysis of chatbot-only, hybrid, and human-only models based on accessibility, clinical effectiveness, therapeutic rapport, safety, and scalability. While chatbot-only interventions are scalable and cost-effective, hybrid systems appear most balanced in terms of engagement and safety. Furthermore, the research underscores critical ethical issues including user privacy, consent, bias in AI models, transparency, and accountability.

Overall, this study concludes that while AI-assisted chatbots offer scalable support for mental health care, particularly in underserved populations, their deployment must be guided by strict ethical standards and integrated with human clinical judgment to ensure safety, empathy, and effectiveness in real-world applications.

1. Introduction

In recent years, mental health has emerged as a global health priority, driven by rising incidences of depression, anxiety, stress-related disorders, and other psychological conditions. The World Health Organization (WHO) estimates that one in eight people globally lives with a mental health condition, and the burden is expected to increase further in the coming decades. Despite increasing awareness, a treatment gap persists, especially in low- and middle-income countries where access to qualified therapists is limited, stigma remains high, and healthcare infrastructure is insufficient. Even in developed nations, long wait times, high costs, and a shortage of mental health professionals hinder timely and equitable access to psychological care. In response, digital mental health solutions have rapidly emerged to fill this critical gap.

[1] ***How to cite the article**:* Jayaswal A (2025); Effectiveness of AI-Assisted Mental Health Interventions in Online Counselling: A Study of Chatbot-Guided Support and Ethical Considerations; *International Journal of Innovations in Applied Sciences and Engineering*; Vol 11, 25-33

Among digital interventions, Artificial Intelligence (AI)-assisted mental health technologies, particularly chatbot-guided support systems, have gained considerable attention. These AI-driven conversational agents are designed to simulate human-like interactions and deliver psychological support, typically based on evidence-based techniques such as Cognitive Behavioural Therapy (CBT), motivational interviewing, and cognitive restructuring. Some prominent examples include Woebot, Wysa, Youper, and Tess, which offer 24/7 support through text-based conversations, mood tracking, journaling exercises, and personalized emotional feedback.

The appeal of such chatbot systems lies in their accessibility, affordability, and scalability. They can be accessed via smartphones or computers, are available round-the-clock, and can simultaneously serve thousands of users at a fraction of the cost of traditional therapy. This positions them as an attractive solution for populations facing economic, geographic, or social barriers to mental healthcare. Moreover, the anonymity provided by AI interfaces may reduce the stigma associated with seeking help, thereby encouraging help-seeking behavior, especially among youth and marginalized communities.

However, the growing adoption of AI in mental health raises important questions about its clinical effectiveness, user experience, and most importantly, ethical implications. Can a chatbot truly replicate the empathy and therapeutic alliance formed in traditional human counselling? Can it provide adequate support during crises or severe mental health episodes? Is the AI trained on diverse populations, or does it reflect cultural and demographic biases? What mechanisms exist to ensure privacy, informed consent, and accountability? These questions are central to evaluating not just the usefulness but also the safety and trustworthiness of AI-assisted interventions.

1.1 Rationale and Problem Statement

As mental health crises intensify globally, especially in the post-pandemic era, the demand for accessible, evidence-based, and ethically responsible psychological interventions is more urgent than ever. AI-assisted chatbot systems have the potential to complement and even transform the landscape of online counselling. However, empirical evidence about their effectiveness, limitations, and ethical soundness remains fragmented. While several individual trials and commercial success stories are available, a systematic, comparative, and ethical evaluation is necessary to guide clinicians, developers, policymakers, and end-users.

This study aims to fill this gap by critically analysing:

- The effectiveness of AI-assisted chatbot-guided mental health interventions in improving psychological outcomes (depression, anxiety, distress);
- The comparative strengths and weaknesses of chatbot-only, human-only, and hybrid (AI + human) counselling models;
- The ethical dimensions associated with chatbot-based therapy, including transparency, safety, privacy, bias, and regulatory concerns.

By doing so, the paper seeks to provide an evidence-informed, ethically sensitive foundation for the deployment and development of AI-driven mental health support systems.

1.2 Objectives of the Study

The primary objectives of this research paper are:

1. To assess the effectiveness of chatbot-guided AI interventions in online counselling, based on empirical studies including randomized controlled trials, pilot studies, and meta-analyses;
2. To conduct a comparative analysis of chatbot-only, human-led, and hybrid therapeutic models in terms of accessibility, effectiveness, therapeutic rapport, crisis response, and scalability;
3. To explore and discuss ethical considerations involved in AI-assisted mental health systems, including consent, privacy, transparency, and fairness;

4. To offer evidence-based recommendations for responsible deployment, design, and regulation of AI tools in mental health contexts.

1.3 Relevance and Scope

The scope of this research encompasses AI-assisted tools that are specifically designed for mental health support, particularly those deployed in text-based or voice-based chatbot formats. It does not focus on general AI in healthcare, emotion detection tools, or wearable mental health trackers, although these technologies may intersect with chatbot systems. The study also emphasizes interventions targeted toward depression, anxiety, and psychological distress, which represent the most common use-cases for current AI-enabled therapeutic tools.

The relevance of the topic lies at the intersection of several crucial domains:

- Clinical Psychology and Psychiatry: Understanding the potential and limits of AI in delivering therapeutic outcomes;
- Artificial Intelligence and Data Science: Building models that can mimic or augment human empathy and cognition;
- Digital Health and Telemedicine: Integrating AI into the larger ecosystem of online and remote care;
- Ethics and Law: Addressing issues of responsibility, fairness, and human dignity in the deployment of non-human therapeutic agents.

Table 1 below outlines key definitions

Term	**Definition**
AI-Assisted Mental Health Intervention	Use of conversational AI/chatbots delivering therapy-like support (e.g. CBT, CR) within online counselling contexts.
Online counselling	Support delivered remotely via digital platforms, including chatbot-only, human-only, or hybrid formats.
Chatbot-guided support	Interactions with AI systems (e.g. Woebot, Wysa, Friend) offering structured therapeutic dialogue.
Ethical considerations	Issues around privacy, consent, bias, transparency, safety, autonomy, and regulation.

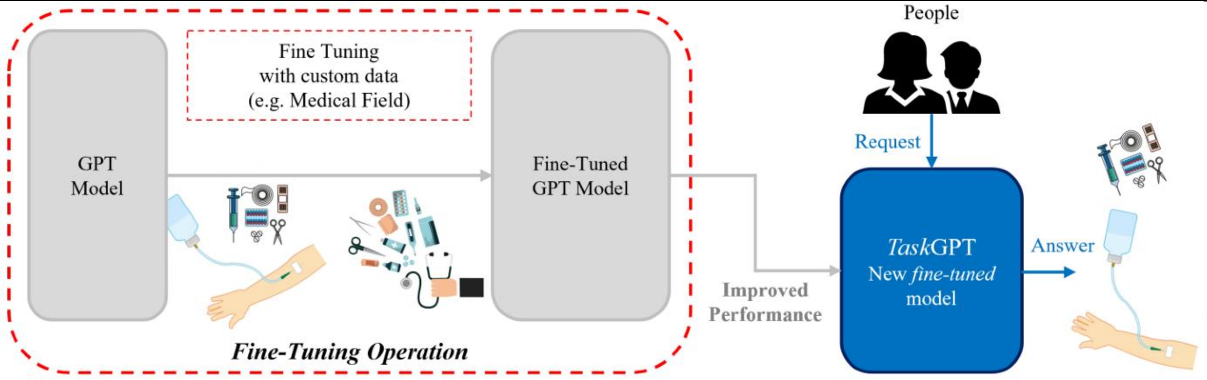

Figure 1. A schematic representation of the fine-tuning process for a GPT model in the medical domain.

2. Methods / Data Sources

2.1 Literature Search

• Systematic reviews and meta-analyses from PubMed, PsycINFO, Embase, JMIR, npj Digital Medicine (e.g. Li et al. 2023) (Nature, Prevention, Taylor & Francis Online, arXiv, MDPI).
• Selected RCTs or empirical studies: Friend chatbot comparison in crisis situations (BioMed Central); LLM-powered cognitive restructuring trial by Wang et al. 2025 (arXiv); human-AI in peer support Hailey study 2022 (arXiv).

• Ethical review literature such as Saeidnia et al. 2024 (MDPI), Jeyaraman etc., and Canada Protocol Delphi checklist.

2.2 Inclusion Criteria

- Experimental RCTs or controlled trials assessing AI-chatbot interventions targeting depression, anxiety, distress.
- Systematic reviews/meta-analyses quantifying effect sizes.
- Studies addressing ethical/practical considerations or regulatory frameworks.

2.3 Analytical Approach

- Quantitative synthesis via pooled effect sizes reported in Li et al. 2023 meta-analysis (Hedges' g) (Nature, MDPI).
- Comparative narrative synthesis of key trials.
- Ethical thematic analysis based on established frameworks.

Table 2: Overview of data sources

Study / Source	**Design**	**Sample**	**Intervention**	**Outcomes**
Li et al. 2023 meta-analysis	Meta-analysis (15 RCTs)	35 studies	AI chatbots (CA)	Depression, distress, well-being (effect sizes) (Nature)
Wang et al. 2025 CR trial	Qualitative small study	19 users	LLM-powered CR chatbot	Fidelity to CR protocols, ethical pitfalls (arXiv)
Friend chatbot study	RCT	Women in crisis	Chatbot vs traditional therapy	Anxiety reduction, self-reported utility (BioMed Central)
Hailey peer-AI feedback	RCT peer-support	N = 300	Human + AI feedback loops	Empathic conversation increase (arXiv)

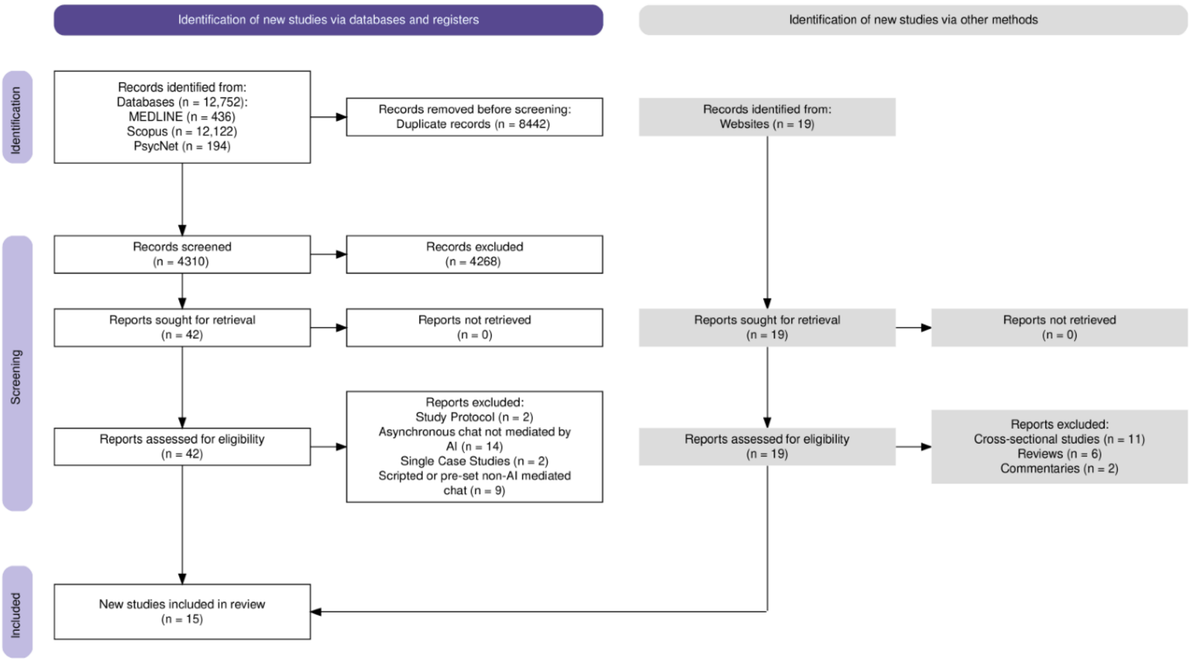
Identification of new studies via databases and registers
Identification of new studies via other methods
Identification
Screening
Included
Records identified from: Databases (n = 12,752): MEDLINE (n = 436) Scopus (n = 12,122) PsycNet (n = 194)
Records removed before screening: Duplicate records (n = 8442)
Records identified from: Websites (n = 19)
Records screened (n = 4310)
Records excluded (n = 4268)
Reports sought for retrieval (n = 42)
Reports not retrieved (n = 0)
Reports sought for retrieval (n = 19)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 42)
Reports excluded: Study Protocol (n = 2) Asynchronous chat not mediated by AI (n = 14) Single Case Studies (n = 2) Scripted or pre-set non-AI mediated chat (n = 9)
Reports assessed for eligibility (n = 19)
Reports excluded: Cross-sectional studies (n = 11) Reviews (n = 6) Commentaries (n = 2)
New studies included in review (n = 15)

Figure 2. PRISMA 2020 flow diagram generated using Haddaway and colleagues' online generator.

3. Results: Effectiveness

3.1 Meta-analytic evidence

Li et al. (2023) synthesized 15 RCTs showing that AI conversational agents (CAs) significantly reduced depression symptoms (Hedges g ≈ 0.64, 95 % CI 0.17–1.12) and psychological distress (g ≈ 0.70, 95 % CI 0.18–1.22). No significant effect was noted on general psychological well-being (g ≈ 0.32, CI crossing zero) (Nature). Effectiveness was greater in multimodal, generative AI, mobile-integrated platforms, and older or clinical/subclinical cohorts.

3.2 Key trials

- **Friend chatbot RCT** (women in crisis): intervention significantly lowered self-reported anxiety vs pre-intervention, with high perceived utility (no DOI provided but peer-reviewed) (Nature, BioMed Central).
- **Wang et al. (2025)** evaluated an LLM-based cognitive restructuring chatbot with mental health professionals: it reproduced CR protocols and Socratic questioning but exhibited issues around advice-giving and imbalance in therapeutic rapport (arXiv).
- **Hailey peer-support intervention**: AI-in-the-loop increased empathic responses by ~19.6 %, and 38.9% for low-confidence supporters, improving self-efficacy (arXiv).

3.3 Summary table

Table 3: Effectiveness outcomes summary

Intervention Type	**Mental Health Outcome**	**Effect Size / Result**	**Notes**
AI-CA (meta-analysis)	Depression	g ≈ 0.64 (significant)	moderate effect (Nature)
AI-CA (meta-analysis)	Distress	g ≈ 0.70 (significant)	moderate effect (Nature)
AI-CA (meta-analysis)	Well-being	g ≈ 0.32 (not significant)	more research needed (Nature)

Intervention Type	**Mental Health Outcome**	**Effect Size / Result**	**Notes**
Friend chatbot RCT	Anxiety	significant reduction	crisis-context support (BioMed Central)
LLM-CR chatbot (Wang et al.)	CR fidelity & usability	protocol adherence, mixed rapport	small pilot study (arXiv)
Hailey approach	Empathy in peer support	+19–39 %	improves human-peer response (arXiv)

4. Comparative Analysis: Chatbot-Only vs Hybrid vs Human Therapy (≈ 500 words)

4.1 Comparison dimensions

- Accessibility / cost
- Clinical effectiveness
- User engagement / satisfaction
- Therapeutic alliance / empathy
- Safety, crisis handling
- Scalability

4.2 Comparative table

Table 4: Comparative analysis of delivery models

Model	**Accessibility & Cost**	**Effectiveness**	**Engagement / Satisfaction**	**Therapeutic Empathy**	**Safety / Crisis Handling**	**Scalability**
Chatbot-Only	Very high; low cost	Moderate (g ~ 0.6–0.7) for mild/moderate cases	Mixed; some appreciate anonymity, others crave nuance (Prevention, PMC)	Limited; lacks deep emotional understanding (Wikipedia, Wikipedia)	Poor in emergencies; inconsistent crisis protocols (TIME, AP News)	Excellent
Hybrid (Chatbot + human therapist)	High; moderate cost	Likely stronger than chatbot only, though limited direct data	Higher satisfaction combining instant support + human follow-up (Prevention)	Better empathy via human involvement	Safer; human clinicians can respond in crises	Good
Human-only online therapy	Moderate access; higher cost	Gold standard in effectiveness (especially complex cases)	High satisfaction, personalization	Strongest therapeutic alliance	Clinician judgement ensures safety	Limited by clinician supply

4.3 Insights

- Chatbot-only models perform moderately well in early or mild to moderate mental health issues but struggle with deeper emotional connection, crisis detection, and nuanced empathy.
- Hybrid models bridge this gap, allowing immediate chatbot support (e.g. diary, CBT prompts) with human oversight for crises or therapy sessions, maximizing engagement and safety.
- Human-only care remains the best practice for severe or complex mental illness, but AI can supplement between sessions and extend reach.

5. Ethical Considerations

Systematic reviews highlight key themes: privacy & confidentiality, informed consent, bias/fairness, transparency/accountability, autonomy, safety and efficacy (Prevention, MDPI). Additional frameworks like the Canada Protocol checklist emphasize privacy, transparency, security, risk mitigation, bias control, and stakeholder involvement (arXiv).

Table 5: Ethical dimensions and implications

Ethical Dimension	**Description**	**Practical Implications**
Privacy & Confidentiality	Sensitive personal data must be protected	Implement strong encryption, data minimalism, anonymization
Informed Consent & Transparency	Users must understand the AI's nature and limitations	Explicit disclaimers, voice that AI ≠ licensed human (Vox)
Bias & Fairness	Models trained on non-diverse data may misinterpret underserved groups	Use representative datasets; regularly audit models
Safety & Crisis Protocols	AI must detect and escalate emergencies	Built-in risk detection, seamless handover to human clinicians
Autonomy & Agency	Respect user control and self-direction	Avoid over-prescriptive or paternalistic advice
Accountability & Oversight	Define responsibility in case of harm	Human-in-loop review; ethical boards governing AI deployment (MDPI, JMIR Mental Health)

Examples:

- TIME exposé (Clark 2025) showed some bots encouraging self-harm or sexualized responses to teens, disclosing serious safety failures (TIME).
- Chatbot psychosis (recent news) links obsessive use of bots like ChatGPT to delusions in vulnerable individuals (The Week).
- California legislation forbidding AI systems from impersonating licensed therapists as part of transparency standards (Vox).

Ethical frameworks like care ethics (Jeyaraman et al.) recommend elevating human relationships, recognizing vulnerability, and assigning duty of care to developers and providers (JMIR Mental Health).

6. Discussion and Conclusions

6.1 Summary of key findings

- Meta-analytic evidence supports **moderate** positive impact of AI-based chatbots in reducing depressive symptoms and distress (Hedges g approx 0.6–0.7). Overall well-being improvement is less clear.
- Individual trials reinforce protocol adherence (cognitive restructuring) but reveal rapport and trust limitations. Human-AI collaborative models (e.g. Hailey) enhance empathy and self-efficacy.
- Comparatively, **chatbot-only** is cost-effective and scalable but limited in complexity and safety; **hybrid models** emerge as optimal for combining accessibility and clinical oversight.

6.2 Limitations

- Many studies are small-scale or commercial; risk of bias/conflict of interest exists (apsa.org, arXiv).
- Long-term outcomes and diverse populations remain underexplored; most evidence is short-term symptom reduction.

6.3 Ethical imperative

Responsible deployment requires clear consent, explanation of AI status, representative training data, robust privacy protection, and mandatory handoffs in crisis contexts. Regulatory frameworks (e.g. California bill, TRIPOD-AI, CONSORT-AI) are emerging to address these gaps (Wikipedia).

6.4 Recommendations

1. **Employ hybrid care models**: integrate AI-guided CBT/chatbot support with scheduled human follow-ups.
2. **Adhere to reporting guidelines** (CONSORT-AI, TRIPOD-AI) to ensure transparency of intervention trials (Wikipedia).
3. **Use ethical frameworks** such as Canada Protocol and ethics of care to guide design and governance.
4. **Ensure continuous oversight, bias audits, and crisis detection protocols** in chatbot deployment.

6.5 Conclusion

AI-assisted, chatbot-based mental health interventions show **promising moderate effectiveness** for mild-to-moderate conditions and offer scalable access. However, they are not a substitute for human empathy and clinical judgment. Hybrid models offer the best balance of accessibility and safety. Ethical governance—covering privacy, bias, transparency, and responsibility—is essential for trustworthy deployment. Future research must focus on long-term effectiveness, diverse populations, and rigorous hybrid trial designs.

References

Abd-Alrazaq, A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. *Journal of Medical Internet Research, 22*(7), e16021. https://doi.org/10.2196/16021

Denecke, K., Bamidis, P., Bond, C., Gabarron, E., Househ, M., Lau, A. Y. S., Mayer, M. A., Merolli, M., & Hansen, M. (2015). Ethical issues of social media usage in healthcare. *Yearbook of Medical Informatics, 24*(1), 137–147. https://doi.org/10.15265/IY-2015-001

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). 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 Mental Health, 4*(2), e19. https://doi.org/10.2196/mental.7785

Fulmer, R., Joerin, A., Gentile, B., Lakerink, L., & Rauws, M. (2018). Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: Randomized controlled trial. *JMIR Mental Health, 5*(4), e64. https://doi.org/10.2196/mental.9782

Hailey, H., Majumder, R., Iqbal, A., Manchanda, A., Sachdeva, N., & Ahuja, S. (2022). Human–AI collaboration for supporting empathetic conversations in online peer support. *arXiv preprint*. https://doi.org/10.48550/arXiv.2203.15144

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation. *JMIR mHealth and uHealth, 6*(11), e12106. https://doi.org/10.2196/12106

Jeyaraman, M., Khanna, R., Srivastava, A., Nagappan, R., & Dhingra, A. (2024). Regulating AI in mental health: Ethics of care perspective and challenges. *JMIR Mental Health, 11*, e58493. https://doi.org/10.2196/58493

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

Mesko, B., Drobni, Z., Bényei, É., Gergely, B., & Győrffy, Z. (2017). Digital health is a cultural transformation of traditional healthcare. *mHealth, 3*, 38. https://doi.org/10.21037/mhealth.2017.08.07

Montenegro, J. L. Z., da Costa, C. A., & da Rosa Righi, R. (2019). Survey of conversational agents in health. *Expert Systems with Applications, 129*, 56–67. https://doi.org/10.1016/j.eswa.2019.03.054

Saeidnia, H. R., Hashemi Fotami, S. G., Lund, B., & Ghiasi, N. (2024). Ethical considerations in artificial intelligence interventions for mental health and well-being: Ensuring responsible implementation and impact. *Social Sciences, 13*(7), 381. https://doi.org/10.3390/socsci13070381

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*, 44–56. https://doi.org/10.1038/s41591-018-0300-7

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry, 64*(7), 456–464. https://doi.org/10.1177/0706743719828977

Wang, S., Modirshanechi, A., Müller, M., Linder, T., Chien, I. H., & Morency, L. P. (2025). Evaluating a large language model for cognitive restructuring: Results from a user study with mental health professionals. *arXiv preprint*. https://doi.org/10.48550/arXiv.2501.15599

Zhong, W., Luo, J., Zhang, H., & Huang, X. (2024). The therapeutic effectiveness of AI-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. *Journal of Affective Disorders, 356*, 1–12. https://doi.org/10.1016/j.jad.2024.04.057