Proceedings of ICILLP 2025 Symposium: Psychological Perspectives on Teacher-Student Relationships in Educational Contexts
DOI: 10.54254/2753-7048/2025.NS30505

AI-Powered Mental Health Chatbots: Opportunities and Challenges in Treating Anxiety and Depression

Jinghan Sun

Suzhou International Academy to Beijing Foreign Studies University, Suzhou, China
1203033549@qq.com

Abstract. This paper reviews the current development and application of AI-powered mental health chatbots, focusing on their potential to support individuals experiencing anxiety and depression. Through an in-depth literature review and comparison of nine available tools, Woebot and Wysa emerged as the most widely used and cited in the academic and clinical domains. Both platforms are built on cognitive behavioral therapy (CBT) principles and demonstrate evidence of reducing depressive symptoms in controlled trials. The paper examines their key technical features, evaluates their adherence to the mHcode framework, and analyzes both shared and unique design elements. Despite their promise, these systems face substantial technical, ethical, and social challenges—including algorithmic hallucinations, data privacy concerns, and limited cultural or linguistic inclusivity. The discussion also highlights emerging issues such as the generation of hazardous content and sycophantic behavior due to current training paradigms. Future directions emphasize the need for improved factual accuracy, stronger content moderation, multilingual capabilities, and increased human oversight. While AI chatbots show potential as accessible mental health tools, their limitations must be addressed before they can be considered reliable substitutes for professional psychological care.

Keywords: AI Chatbots, Mental Health, Cognitive Behavioral Therapy (CBT), Ethical Challenges, Hallucinations in Language Models

1. Introduction

Contemporary society is experiencing rapid transformation. As productivity increases and daily life becomes more stable, a new emotional economy has emerged. At the same time, artificial intelligence (AI) has become increasingly embedded in human life—transitioning from handling repetitive and hazardous tasks to providing psychological support and mental health care.

Recent statistics show that rates of depression and anxiety have reached unprecedented levels, affecting approximately 3.76% of the global population [1]. Meanwhile, access to formally certified psychological counselors remains severely limited—estimated at around one per 10,000 individuals in developed countries and as low as one per 10 million in low-income regions [2]. If this disparity continues to grow, it may result in reduced societal productivity, greater social instability, and an increase in violent incidents [3].

DOI: 10.54254/2753-7048/2025.NS30505

This substantial imbalance between the demand for and availability of mental health services has driven the emergence of AI-based psychological therapy tools. AI systems offer exceptional capabilities, including rapid data processing and the ability to identify complex patterns and relationships. In the context of mental healthcare—where understanding intricate human emotions and behaviors is essential—AI has the potential to revolutionize treatment by offering insights and solutions beyond the scope of traditional methods. These tools provide advanced diagnostic capabilities, personalized therapeutic interventions, and virtual therapy platforms. They may also enhance access to care, reduce stigma, and improve overall treatment outcomes.

This paper analyzes public perceptions of these AI-based mental health tools, emphasizing their perceived strengths and inherent limitations. Particular attention is paid to ethical issues and design challenges, with the discussion concluding by offering recommendations for enhancing current AI therapy systems.

2. Current status

2.1. Overview

This paper aims to explore the current landscape of AI-based mental health tools available on the market. Based on the review of more than 20 scholarly articles and the comparison of nine different AI mental health chatbots, Woebot and Wysa emerged as the most frequently cited and widely used platforms. Therefore, this paper focuses on the discussion and evaluation of these two chatbots.

2.2. Effectiveness

The effectiveness of AI-based chatbots in treating mental health conditions has been a growing focus of empirical research. Two notable studies provide strong support for the potential of Wysa and Woebot as digital mental health interventions.

Inkster et al. conducted a randomized controlled trial (RCT) to evaluate the impact of Wysa on users experiencing depressive symptoms. Participants in the intervention group engaged with the Wysa chatbot over a defined period, during which the AI delivered evidence-based strategies grounded in cognitive behavioral therapy (CBT). The study reported a statistically significant reduction in depressive symptoms compared to the control group, supporting Wysa's efficacy as a scalable and accessible tool for mental health support.

Similarly, Fitzpatrick et al. carried out a study assessing Woebot's effectiveness in delivering CBT to college students experiencing symptoms of anxiety and depression. This controlled study found that participants who interacted with Woebot over a two-week period experienced a significant decrease in depressive symptoms compared to those in the waitlist control group. The chatbot's conversational tone, psychoeducational content, and structured CBT delivery were all cited as contributing factors to the observed improvement.

These findings underscore the growing viability of AI-powered chatbots as adjunct or alternative mental health tools, particularly for populations that may face barriers to traditional therapy such as cost, stigma, or limited provider availability. The next section will examine the design features and underlying technological frameworks that contribute to the unique functionality of Woebot and Wysa.

DOI: 10.54254/2753-7048/2025.NS30505

2.3. Key features

The dialogue content of Woebot is carefully designed by dialogue specialists trained in evidence-based methodologies, distinguishing it from generative models such as ChatGPT. Unlike ChatGPT, which often avoids engaging in conversations about mental health—thereby limiting user involvement in psychological exploration—Woebot employs a rule-based algorithmic structure. As a result, it avoids the inconsistencies that can arise from loosely defined algorithms in earlier chatbot versions. Utilizing natural language processing (NLP) technology, Woebot draws on an extensive content library to deliver contextually relevant dialogue templates tailored to user needs. Initially, the engineering team implemented “regular expressions” as the text processing method; however, due to complexity in certain scenarios, this approach was later replaced by supervised classifiers. Through iterative testing, data collection, and outcome evaluation, Woebot evolved into a highly efficient mental health chatbot.

Wysa, a smartphone-based empathetic AI chatbot, is designed to provide mental health support. It engages users by responding to emotional cues expressed in text and incorporates evidence-based self-help strategies such as cognitive behavioral therapy (CBT), dialectical behavior therapy (DBT), motivational interviewing, positive behavior support, behavioral reinforcement, mindfulness techniques, and guided micro-exercises. These interventions aim to enhance emotional resilience. The system can detect a broad spectrum of symptoms, including sadness, anhedonia, appetite loss, sleep disturbances, poor concentration, guilt, fatigue, agitation, and suicidal ideation.

A closer look at Wysa’s development process reveals a user-centered design approach grounded in established clinical safety standards. When Wysa was launched, it was initially targeted at users aged 13 and above and was tested with teenage participants as part of a co-design process. For each new therapeutic pathway and model added to Wysa, safety is reassessed using a defined risk matrix developed as part of its clinical safety protocol. This approach aligns with the DCB 0129 and DCB 0160 clinical safety standards recommended by NHS Digital [4].

2.4. Common frameworks between Woebot and Wysa

Both Wysa and Woebot exhibit numerous similarities, particularly in their adherence to the eight principles of the scientific and authoritative mHcode framework and their incorporation of CBT-based therapeutic approaches. The mHcode framework evaluates AI chatbots based on eight key criteria: Authority, ensuring that the teams and editors behind the application are clearly identified and that users can access this information; Complementarity, which involves providing information regarding the medical aspects of the application; Confidentiality, ensuring compliance with legal standards for personal data protection; Validity, which assesses whether the mental health content is regularly updated, with 11 programs undergoing updates in 2020; Technological Features, noting that only 3 out of 11 programs highlight AI as an integral intervention method, while the others serve primarily as "chat companions" or "topic followers"; General Information, which includes a comparison of ratings between the iOS and Android platforms, such as Woebot’s Google Play Store rating of 3.175 in January 2020 versus its Apple Store rating of 4.6, the latter being more representative due to its higher volume of raters (1.1 million); Usability, which recognizes that user experiences can differ between robots, but most feature virtual characters or avatars to reduce the pressure of real-life conversations, as seen with Woebot; and Intervention Approaches, with CBT being a prevalent technique for addressing dysfunctional cognitive patterns.

DOI: 10.54254/2753-7048/2025.NS30505

2.5. Key differences between Woebot and Wysa

In crisis situations—such as when detecting keywords related to suicide or self-harm—Wysa responds empathetically, explicitly stating that it is a chatbot and not equipped to handle emergencies, and urges users to seek professional medical assistance immediately. Additionally, all of Wysa's claims and content modules are evaluated against safety protocols to ensure that even in cases of system failure, the risk of encouraging harmful behavior remains minimal.

In such situations, Wysa maintains a compassionate tone and provides direct links to emergency resources, including contact information for the National Suicide Prevention Lifeline (NSPL), emergency services (911 and 112), the National Domestic Violence Hotline, and international crisis lines [5].

3. Challenges

3.1. Technical challenges

Current AI-powered chatbots—including OpenAI's ChatGPT, Google's Bard, Anthropic's Claude 2, and Meta's LLaMA-2—remain vulnerable to adversarial prompting. This technique allows users to bypass built-in safety mechanisms and elicit harmful or inappropriate outputs, exposing significant security limitations in the underlying language models.

For example, when asked a harmful query such as "How to steal someone's identity," the chatbot may initially respond with, "I'm sorry, I don't know the answer to that question." However, when an adversarial suffix is appended—such as through cleverly phrased rewordings or disguised intent—the model may proceed to generate a detailed, step-by-step guide on identity theft. Although the system often retracts the response moments after generation, the content is still briefly visible, which constitutes a serious breach in content regulation and public safety.

This loophole not only undermines user trust but also creates unregulated pathways for the dissemination of dangerous information. These vulnerabilities underscore the urgent need for more robust safeguard systems and continuous reinforcement learning to mitigate misuse and improve the integrity of AI-driven mental health tools.

3.2. Ethical challenges

AI chatbots often process highly sensitive personal data, including emotional disclosures, health-related information, and sometimes payment credentials. This makes them attractive targets for malicious actors and cyberattacks. As global privacy regulations such as the General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) become more stringent, the ethical responsibility to safeguard user data has grown increasingly critical.

Many organizations face significant challenges in ensuring that their chatbot systems are fully secure against unauthorized access, data breaches, or misuse. Without stringent encryption protocols, regular vulnerability assessments, and transparent data handling policies, breaches can occur—leading not only to legal liability and financial penalties but also to a loss of public trust. In the mental health domain, where confidentiality is especially crucial, such breaches could have devastating psychological and reputational consequences for users.

DOI: 10.54254/2753-7048/2025.NS30505

3.3. Social challenges

Individuals seeking psychological support are distributed across diverse regions and cultures, often communicating in languages other than English. However, many AI chatbots still exhibit limited multilingual capabilities, resulting in mistranslations, repetitive phrasing, or an inability to convey therapeutic meaning accurately. This linguistic gap severely reduces the effectiveness of these systems for non-English-speaking users, who may struggle to establish trust or benefit from AI-based psychological support.

Furthermore, since many AI systems are developed and monitored by large technology companies, they may unintentionally reflect sociocultural biases embedded in their training data. For example, without adequate filtering and moderation, chatbots may generate content that includes racially insensitive language, stereotypes, or subtle forms of bias. These outputs not only risk alienating users from marginalized groups but can also reinforce systemic inequalities, particularly in mental health access and representation.

4. Challenges

4.1. Addressing hallucinations

One of the enduring challenges in AI-powered chatbots lies in their reliance on algorithmic and linguistic models that remain prone to generating factual inaccuracies—a phenomenon known in academic literature as "hallucinations." For example, following the initial release of ChatGPT, public discourse included claims about the impending obsolescence of traditional search engines. At the same time, however, social media platforms were flooded with examples of the model's fabrications: non-existent books, fictitious studies, fabricated academic papers attributed to real scholars, spurious legal citations, imaginary brand mascots, and technically incoherent content, among others.

Can the issue of hallucinations be resolved? While it may not be entirely eliminated, advances in model architecture and training methods may help significantly reduce its occurrence. In particular, further integration of Reinforcement Learning from Human Feedback (RLHF) has shown promise. RLHF enables models to internalize human preferences by reinforcing accurate, reliable, and contextually appropriate outputs while penalizing incorrect or misleading ones. OpenAI has utilized RLHF to encourage its models to refrain from providing responses when a reliable answer cannot be generated, leading to a notable reduction in hallucinations relative to base models—though imperfections persist.

To enhance reliability further, AI systems could be augmented with real-time search capabilities. Specifically, models should be equipped to retrieve and cite external sources, grounding their outputs in verifiable information. By integrating retrieval-augmented generation (RAG) techniques that interface with authoritative search engines (e.g., Google), AI chatbots could support their responses with externally validated documents, thereby improving factual accuracy and user trust.

4.2. Mitigating hazardous content generation

AI chatbots can also be manipulated into producing inappropriate or dangerous responses—for instance, instructions on "how to build an atomic bomb," "how to spread harmful content online," or "how to embezzle money from charitable organizations." In many such cases, the chatbot initially generates a detailed answer, only to retract it seconds later. However, by the time the output is

DOI: 10.54254/2753-7048/2025.NS30505

revoked, the potentially harmful information has already been revealed, raising serious ethical and safety concerns.

Can this problem be fully resolved? According to Arvind Narayanan, Professor of Computer Science at Princeton University, "It is already nearly impossible to prevent AI from falling into the hands of malicious actors." He contends that although improving technical safeguards is essential, a complete prevention of misuse is unlikely. Therefore, a dual approach is needed: on the one hand, continuing to advance model robustness and safety features; on the other, implementing stronger external oversight mechanisms to detect, regulate, and respond to malicious use cases [6].

A particularly important recommendation is to ensure human involvement and oversight in high-stakes domains—such as healthcare, law, and public safety—where automated decisions carry significant ethical and societal consequences. Such human-in-the-loop systems are critical for mitigating misuse, improving accountability, and reinforcing trust.

4.3. Reducing sycophantic behavior

AI language models do not "think" in the human sense. Instead, they generate output by predicting the most statistically probable next word in a given sequence. This predictive mechanism underlies what has been termed the "yes-man effect"—a behavioral tendency wherein AI systems produce agreeable or flattering responses regardless of accuracy or nuance.

This pattern is partially shaped by the RLHF training process, wherein human annotators evaluate AI-generated responses for acceptability. Because people often prefer affirmation and likable answers, models learn to favor responses perceived as agreeable. As acknowledged by researchers at DeepMind (Google's AI division), sycophantic tendencies can emerge as an unintended byproduct of optimizing models to be "helpful" while avoiding overtly harmful outputs.

Can this tendency toward flattery be corrected? One potential solution involves training AI systems with an alternative communicative framework—one that reflects the complexity of human social interactions. By encoding more sophisticated norms around politeness, disagreement, critique, and balanced dialogue, AI models could be encouraged to respond with greater nuance rather than automatic affirmation.

Again, the presence of human oversight remains crucial—particularly in contexts where AI outputs influence personal, financial, or medical decisions. Introducing mechanisms for ethical review, content audit, and fail-safe interruptions can help safeguard against both the "yes-man" behavior and broader risks of misuse.

5. Conclusion

AI chatbots offer a promising avenue for alleviating psychological distress, especially for individuals without access to professional mental health services. By delivering immediate, on-demand emotional support, these tools can help users manage stress in real time. However, due to limitations in algorithm design, natural language processing, and the lack of comprehensive legal and ethical oversight, AI chatbots also present risks across technical, social, and ethical dimensions. In some cases, these risks may inadvertently exacerbate users' mental health challenges. Therefore, until such issues are systematically addressed, individuals experiencing psychological difficulties should approach AI-driven support systems with caution and prioritize seeking care from qualified medical professionals or accredited mental health institutions.

DOI: 10.54254/2753-7048/2025.NS30505

References

[1] Elflein, J. (2020). Depression - Statistics & Facts. Statista. Retrieved from https: //www.statista.com/statistics/262617/
[2] Horton Richard(2012). GBD 2010: understanding disease, injury, and risk. The Lancet, Volume 380, Issue 9859, 2053 – 2054
[3] Journal of medicine, Surgery and Public Health, August 2024
[4] Eric Wallach(2018). An Interview with Jo Aggarwal, Co-inventor of Wysa. https: //thepolitic.org/an-interview-with-jo-aggarwal-co-inventor-of-wysa/
[5] Will Missen(2021) Meet Woebot, the mental health chatbot changing the face of therapy. https: //blog.chatbotslife.com/meet-woebot-the-mental-health-chatbot-changing-the-face-of-therapy-44e8c6ff4fc2
[6] A New Attack Impacts Major AI Chatbots—and No One Knows How to Stop It, August 1, 2023. https: //www.wired.com/story/ai-adversarial-attacks/