

Effectiveness of AI-Based Conversational Agents in Mental Health Interventions: A Meta-Analysis

Adyasha Samal¹, Pavithra V ^{*}2, Uma Shankari Srinivasan ³ R.
Renuga Devi ⁴

^{1,2,3,4}Department of Computer Science and Applications, Faculty of Science and Humanities, SRM
Institute of Science and Technology, Chennai, Ramapuram -600089, India.

adyashasamal04@gmail.com¹, vpavithra.1989@gmail.com², uma.balaje@srmsit.ac.in³
drrenugadevi86@gmail.com⁴

Abstract

Mental health issues such as stress, depression, anxiety and emotional instability remain to be extensive public health challenges, along with limited access to traditional treatment, especially in low-resource situations. Artificial Intelligence (AI) based chatbots or conversational agents have arisen as cost effective and scalable tools for providing mental health care. This integrative review and meta-analysis synthesize researched from 20 peer-reviewed studies assessing the effectiveness of AI-based chatbots in mental health support. The review encompasses 3 main dimensions: [1] the aggregated effectiveness of chatbots in decreasing symptoms of emotional instability, anxiety, depression, and stress; [2] the role of personalization in improving engagement and therapeutic results; and [3] the relative impact of short-term vs long-term applications. The results indicate that AI chat agents, especially those based on cognitive behavioural therapy (CBT), exhibit reasonable effectiveness in decreasing mental health symptoms, particularly in short-term intervention. Personalization significantly increases both the treatment outcomes and engagement, with higher retention and improvement in symptoms as described by users who received tailor-made interventions. Nevertheless, the literature discloses a significant gap in long-term follow-up data, preventing conclusions about persistent benefits. Although AI conversation agent provides to be effective tools in mental health support, further sequential research is necessary for assessing their long-term clinical usage and integrating into hybrid support models.

Keywords: *AI chatbots, mental health, conversational agents, personalization, meta-analysis, depression, anxiety, digital intervention*

1.Introduction

Mental health disorders have become a growing issue worldwide as World Health Organization (2023) reported that in 2019, over 970 million people suffered from mental conditions such as anxiety, stress and depression. The COVID-19 pandemic made it even worse as the anxiety and depression rates were increased by 25% worldwide. Such conditions will not just affect the wellbeing of individuals, but also their performance in academics, productivity at workplace and overall quality of life (Kuehn, 2022). This rising trend have increased the need for mental health services more than ever but still the health professionals are not sufficient in many under developed and developing countries to provide for low- and middle-income groups (Patel et al., 2018).

Even in the developing countries, many challenging factors like high treatment costs, long waiting periods and stigma related to mental health care and commuting issues[4] that stops them from getting support on time. Due to this, people in millions go without any treatment or even discontinue their care prematurely. These complex problems have encouraged the clinicals and researchers to explore solutions through digital avenues that are easy to access and scalable as well.

This situation has led to the rise of digital mental health avenues that ranges from teletherapy, mobile health applications to online programs. The most recent advancements in the digital mental health support are the chatbots which are technically known as conversational agents powered by AI models proved to be most novel and effective way for delivering mental health support. These agentic chatbots leverages natural language processing (NLP) and advanced machine learnings algorithms for stimulating human dialogue to deliver support based on psychological frameworks like Cognitive Behavioural Therapy (CBT), Psychoeducation, Motivational interviewing and Mindfulness-based strategies (Fitzpatrick et al., 2017; Inkster et al., 2023).

Chatbots overthrows the barriers that are generally posed in traditional therapy such as 24/7 availability, low-cost or cost-free, and anonymity helps to grab to attention of those people that avoids face-to-face consultation. Wysa, Emohaa, Woebot and Tess are some of the popular bots that are known for helping people with depress, stress, loneliness, trauma and anxiety (Suharwardy et al., 2023; Sabour et al., 2023). These chatbots have the[5] capability to provide interactive and structure conversations that can aid its users in managing emotions and thoughts, monitor their moods and inculcate mindfulness.

The initial researches on chatbot agents in mental health support had indicated reasonable decrease in anxiety and depression symptoms, especially on people with lower to medium level conditions. But the results are varied across studies as some results are statically significant while reports show minimal impact[6] when compared with control groups such psychoeducation email and self-help books (Potts et al., 2022; Klos et al., 2021). Especially the effectiveness of chatbots pivots on factors such as interaction frequency, level of personalization's and user engagement.

Moreover, there is noticeable gap on long-term follow-up in the existing literature. The sustainability of the benefits from chatbot agents after initial treatment periods are only covered in few studied while comparison between chat agents and human support[7] interventions are assessed in even fewer studies. The key driver for user retention is mostly considered due to personalization while treatment adherence is mostly unclear or improperly used on various studies (Manole et al., 2024; Inkster et al., 2023). These are the challenges that impacts the curation of best practices for chatbots and utilizing them in mental health care.

In order to bridge these gaps, this systematic review enlightens the findings of 20 peer-reviewed studies that assessed the effectiveness of AI-based chatbot agents for mental health interventions. In particular, this paper seeks for three fundamental analysis questions:

1. How effective are AI chatbots in lessening symptoms of depression, anxiety and dress?
2. Does personalization improve user engagement and treatment outcomes?
3. How do short-term outcomes compare to long-term effectiveness?

2. The Role of Personalization in Engagement & Treatment Outcomes

Personalization plays the key part when it comes to chatbot agent's effectiveness in mental health support. Providing personalized chat agents by modifying response and content based on the user's requirements helps to improve not only the user engagement, but also the end-result of the treatment[8]. From the reviewed studies, we can clearly identify that personalized methods outweigh the non-personalized methods especially when it comes to retaining the users, improving the interactions and provides highly effective treatment outcomes.

2.1 Impact of AI driven Chatbots on Mental Health

2.1.1 Generic impact trends

A repetitive pattern over 20 reviewed studies claims that AI driven communicative agents like Chatbots can adequately lessen the psychological symptoms such as excessive anxiety, nervous breakdown, stress and even depression. Mainly, it is been observed that these AI Chatbots has a foundation on evidence-based frameworks, concretely Cognitive Behavioral Therapy (CBT), which are deployed in multiple platforms like messaging platforms, apps, and many more. Intercede generically include structured psychoeducational context, emotions tracing features, thought reformulating exercises, and mind conscious prompts.

MacNeill et al. (2024) tested a randomized controlled trial where he recorded a four-week period scores of people undergoing through chronic illness as of diabetes or arthritis suffer significant reductions in anxiety (GAD-7) and depression (PHQ-9). These readings predicted that the intervention of Chatbots can elevate an individual psychological health problem that leads to improvised physical health conditions, by facilitating with non-stigmatizing support and scalable options.

Similarly, Generative Pre-trained Transformer powered Chatbot customized with adaptive responses and CBT logic was evaluated by Manole et al. (2024). Reportedly the study analyses a reduction in anxiety symptoms at a rate of 21.15% after one week of daily usage ($t(49) = 10.24, p < .001$)[9]. Particularly, some durability in effects has been observed after the maintenance of records for two months- a critical aspect of mental health interventions where the risk of relapse is high. However, the outcomes reported across studies were not consistently positive.

Many have observed that the Chatbot interventions tend to be most effective when compared to passive or minimal-intervention controls, such as self-help PDFs or waitlists, but they often fall short in achieving[10] the depth and outcomes provided by therapist-led programs.

2.1.2 Subdivision impact by population

Depending on their clinical and demographics profiles the effective outcomes are based, where multiple factors like chronic illness, postpartum status or academic stress plays a vital role in detecting the effectiveness.

2.1.2.1 University students:

It is been observed in Argentina in a pilot RCT carried out by Klos et al.(2021) that many University students(N=181) are engaged with Tess, a Chatbot that uses Facebook Messenger to deliver motivational interviewing techniques and CBT. It is been noticed by experimental group that there is a substantial decrease in anxiety ($t(26) = 2.15, p = .04$) [11], whereas there was no significant change in level of depression ($p = .08$). This result in a potential domain-specific effect in which anxiety is more responsive to quick intervention whereas depression showcases more frequent ingrained symptoms. This result in high dropout rates in both the groups highlighting how difficult it is to keep youngsters engaged in self-guided technologies.

2.1.2.2 Postpartum Women:

Women undergoing through postpartum suffers from many mental health issues which marks an unique challenge for all. As their mental health get affected by sleep disruption, caregiver burden, and stigma. 192 women tested for 6 weeks RCT at Stanford, Suharwardy et al. (2023) within their postpartum window. Its been observed that there is a reductions in depression among women ranging from mild to moderate baseline symptoms. The chatbot predicted significantly greater change in depression level in comparison to control (PHQ-9: $p = .025$). Without further customization the anxiety outcomes may not be board spectrum effective.

2.1.2.3 Chronic Illness Populations:

Traditional theory often doesn't address about the psychosocial burden carried due to chronic disease. Chatbots users with chronic illness suffering from depression and anxiety[12] has a significant decreased in their disorder as shown by MacNeil et al.(2024). Mainly due to its anonymity and availability, the chatbots are highly appreciated by these participants 24/7, which gives a path for in-person therapy.

2.1.2.4 Subclinical Youngsters:

Within two weeks RCT the *Fido* chatbot was compared to a CBT-based self-help book among Polish young adults by Karkosz et al. (2024). It was notice that there was improvement in both the groups of depression and anxiety where chatbots users were able to complete the intervention in less time. Then the follow up was done after one month showed improving effects, showcasing short-term chatbot may be time-effective for low rate mental health needs.

2.1.3 Chatbot vs. other interventions

2.1.3.1. Against Psychoeducation:

Studies show that chatbots are much more effective than minimal or passive interventions. On receiving structured psychoeducational emails , Gutu et al. (2024) evaluated Woebot against them. When delivered consistently high quality written content were not been outperform properly by CBT-based chatbot.

2.1.3.2. Against Self Help Books:

The self-help CBT books offered higher engagement at lower time cost still they were unable to deliver the desire outcomes[13]. Example: While the control group spent 118 minutes with the books meanwhile chatbot users engaged for 79 minutes in total.

2.1.3.3. Against Human Therapy (or Hybrid Models):

Parity with traditional therapy evidence comes from hybrid approaches. In comparable to face-to-face CBT(GAD-7: $d = 1.6$ vs. 1.3) AI chatbot content with clinician check-ins reduced anxiety scores found by Palmer et al. (2024) in a large trial (N=299). This gives a idea for collaboration between AI and human therapist, providing better clinical depth and scalability.

2.2 Personalization boosts higher engagement

Personalized chatbot agents tailor their engagement according to the individual user's input that stimulates them to feel more related and appealing. In the study of Manole et al. (2024), personalized support had resulted in reducing anxiety of 21.15% users within a single week of use and these advancements continued more than 2 months ($t(49) = 10.24, p < .001$). The personalized chat agents that tailored according to the[14] emotional state of the user had led to higher rates of retention, in comparison to non-personalized ones that had higher rate of dropouts. This provides a clear view that the chances of user's engagement and benefits are higher when the chatbots responses are in alignment with the user's need.

The impact of personalization on user engagement is very high In the scenario where the people go through severe mental health problems, similar like the ones with chronic conditions. The study of MacNeill et al. (2024) shows patients with chronic conditions that used a personalized chat agents for four weeks had resulted in significant reduction in both anxiety and depression.

2.3 Personalization Improves Treatment Results

Personalized chat agents are highly effective in reduction of mental health symptoms in comparison to non-personalized methods, particularly for illness like depress and anxiety. In Suharwady et al. (2023), the postpartum women who used Woebot chatbot with personalized content had felt a tremendous decrease in depression when compare to those groups received standard intervention[15]. It was also revealed that patients experienced a huge involvement with chatbot and this might be the reason for their improvements.

On the contrary, the study of Gutu et al. (2021) shows that chatbot with less personalization that provides common psychoeducation had not perform better any better than those control groups who received psychoeducation through daily emails. This study clearly indicates the demerits of non-personalized approaches and indicates the necessity of tailoring the chatbot based on the user psychological and emotional conditions for better results.

In addition to this, the study of Inkster et al. (2023) shows that personalization improved the chat agent's effectiveness, especially with highly engaged users. The users who interacted more often and experienced the personalized content of the chat agent had shown substantial decrease in anxiety and depression. On the flip side, the users who had lower engagement with the chatbot experienced minimal improvements. This result further highlights that a personalized method, along with continuous user engagement is important for maximizing the treatment outcomes.

2.4 Subgroup Analysis: Personalization Across Populations

The efficacy of personalization differs across various user population as demonstrated by several studies in this review.

- **University Students:** the study of Klos et al. (2021) assessed the effectiveness of a chatbot named Tess, which offers customized cognitive behavioural therapy and motivational interviewing to university students. This study showed a substantial reduction in anxiety but no significant decrease in depression level. This study also indicated comparatively high dropout rates (39%) which suggests that personalized responses may help in minimizing anxiety, but it may not be enough to keep engagement for more complicated mental problems like depression.
- **Postpartum Women:** As mentioned previously, Suharwady et al. (2023) study revealed that personalized chatbot had significant impact on reducing the depression of postpartum women with moderate to severe symptoms. On the contrary, personalization did not substantially reduce the anxiety levels of this group which shows that personalization may have more impact for certain conditions but might need further modification for other symptoms.
- **Chronic Illness Groups:** The study by MacNeil et al. (2024) showed that personalized chatbot was specifically beneficial for people with chronic conditions like diabetes or arthritis. Customized interventions were related to significant reduction in both anxiety and depression [16]with individuals in the group showed higher engagement and satisfaction to the personalized intervention compared to the non-personalized ones.
- **Young Adults:** Karkosz et al. (2024) study indicates that a customized CBT chatbot named Fido had substantially reduced the symptoms of anxiety and depression among young adults that have subclinical conditions. The improvements were persistent for about a month and users had less time commitment compared to other control group that uses self-help book. These results are evident that customized chatbots can be effective and appealing option for people with lower-level mental health conditions.

2.4.1 Comparative Analysis: Personalized chatbots vs Other Interventions

Comparison of personalized chat agents with other approaches such as self-help books, human therapy, and psychoeducation, discloses the substantial advantages of personalization.

- Chatbot vs Self-Help Books: Karkosz et al. (2024) study reveals that personalized chat agent (Fido) demonstrated significant results compared to the self-help book *Mind Over Mood*, while chatbot users spent 79 minutes which is noticeably lesser time than self-book users who had to spend 117 minutes [17]. This result shows that personalized chat agents can offer same benefits in reducing symptoms while needing lesser time commitment that will help to retain the users.
- Chatbots vs Psychoeducation: The study by Gutu et al. (2021) revealed that both personalized chatbot (Woebot) and a daily psychoeducation group reduced the user's anxiety while there's substantial difference between the two. This finding indicates that passive tools like psychoeducation email are outperformed by personalized interventions in terms of effectiveness and engagement as well.
- Chatbot vs Human Therapy: The combination of both chatbot and human therapist observation show the most significant results for attaining the effectiveness of clinical-level. The study of Palmer et al. (2024) demonstrated that a hybrid program with combination of chatbots with clinician oversight, led to reduction in anxiety compared to face-to-face. This shows that personalized chat agents may reach the efficacy of traditional therapies when implemented as a part of hybrid model.

2.5 Long-term effects and sustainability of Personalization

Although personalization has demonstrated substantial short-term effectiveness, the long-term impact in understanding the personalized chatbot is still having a gap. The study of Manole et al. (2024) had shown some continuous improvements [18] for two months, however longer follow-up periods are necessary for evaluating the sustainability of these benefits.

Further long term study is necessary for assessing the potential for personalized chat agents for user interventions on long run, particularly in averting relapse and continuous improvements in mental health. By assessing the ongoing personalization's through real-time feedback or predictive analysis, may provide more customized long-term treatment that adheres to changing needs of the users.

2.6 Methodological Consideration in Measuring Personalization

Although personalization is considered as helpful but assessing the impact across studies have challenges. Many research studies often use various definitions of what comes under personalization in AI chatbot agents. It starts from a simple mood monitoring and tailored responses to highly dynamic and sophisticated alterations. Difference in the implementation of personalization (e.g. personalized CBT exercises, user-specific progress) makes it complicated to directly make comparison on different studies.

Moreover, some studies like the Potts et al. (2022) indicates the efficacy of personalized chatbots is highly influenced by high user engagement. But the studies that overlooks on measuring engagement levels or collect user feedbacks may exaggerate the impact of personalization. Future study should make a standardized metrics list for assessing the personalization and engagement to enable more solid comparisons.

3. Short-Term Vs Long Term Effectiveness of AI Chatbots in Mental Health Support

The most crucial question which is yet to be answered on the AI Chatbot intervention in mental health is whether the improvements noticed in short-term usage are maintained in the long-term. Although

many studies highlighted that brief intervention of chatbot agents results in statistically significant reduction in the symptoms of mental health issues, very limited had follow-up tests for evaluating the efficacy on long-term. The studies involved in our review has shown solid pattern in short-term improvements especially in depression and anxiety, however very less evidence of strong long-term improvement, mainly due to the observation window being short mostly.

3.1 Solid Pattern on Short Term Benefits

Most of the studies reviewed had demonstrated results soon after the intervention, mostly starting from 1 to weeks to 6 weeks after using the chatbot. In general, the outcomes showed statistically major reduction in depression, stress or anxiety symptoms.

- In the study of Manole et al. (2024), there was a 1-week intervention with the help of custom-made CBT chatbot agent seen a 21.15% decrease in symptoms of anxiety. This was the quickest and most crucial in the entire review showing the effects in short term.
- The study by MacNeil et al. (2024) involved a 4-weeks treatment with Wysa chatbot among people with chronic conditions and they had showed a significant reduction in GAD-7 and PHQ-9 scores. The developments were noticed in the week four but no follow-up was performed after the trial.
- In Suharwady et al. (2023) study, the users were individuals in postpartum who use Woebot chat agent had revealed strong evidence of reduction of depression symptoms after 6 weeks but follow-up was not performed to observe the continuity.
- In the study of Klos et al. (2021), university students who had chatbot interactions starting to see reduction in anxiety within 2 weeks, however long-term follow up were omitted again.

These studies indicate the interventions by chatbot can result in significant improvements for a short period ranging from 1 to 4 weeks, in particular to anxiety symptoms. But most of the trials conducted in these studies assessed soon after the intervention, leaving the space for the question about their long-term sustainability.

3.2 Less Long-Term Follow Ups

There are very limited studies in the review that conducted follow-ups for assessing the effects outside the post-intervention time, and they were short as well usually 1 or 2 months at maximum.

- In the Manole et al. (2024) study, there was a 2 month follow up conducted after a one-week gap of treatment. The outcomes indicated a decrease in anxiety have sustained in the follow-up assessment shows the strength of the chatbots effect sustaining at least 8 weeks.
- The study of Karkosz et al. (2024) also had a one-month follow-up after the post-intervention in which the individuals using Fido chatbot had seen a continuous reduction in the anxiety and depress scores. The consistency of the impact for a month backs the short term to intermediate benefit in the chatbot usage but still it leaves the room of long-term effectiveness.
- The study by He et al. (2022) showed a significant decrease in the PHQ-9 scores soon after the treatment but the improvements had reduced in the follow-up after one month indicates a potential decline or the necessity for booster session to have sustainable impact.

The lack of data for backing long-term effect is a huge limitation in the present literature which makes it difficult to assess the sustained value provided by the chatbot interventions. In addition to this, lack of consistent standardized assessment tools or intervals on studies makes it difficult to make long term comparison for getting a meaningful result.

3.3 Gaps and Patters in Short-Term vs Long-Term Effectiveness

The comparison of short-term and long-term results shows few crucial patters:

- Chatbot involvement has a rapid impact on anxiety symptoms and have higher chances of showing long term impact in the follow-up assessments.
- Depression symptoms usually show a reduction initially but the impact usually gets stagnant or reduces over time as showed in He et al. (2022)
- Stress-related progress are infrequently reports and usually don't last long as there are environmental and situational factors influencing stress that a chatbot have no control.
- Personalized interventions like in Manole et al. (2024) suggests that it has slightly higher sustainable improvements in comparison with non-personalized ones, supporting the results in previous section.

The lack of studies measuring the results 3 months or longer to post interventions pose a huge gap in the literature. There was no study that had 6 month or 12 months follow-ups as they are usually standard in clinical trials. This outlines a huge area for future research, particularly if the chatbots are given as potential substitute or assistants to long-term treatment.

3.4 Implications for Clinical Application

From a clinical perspective, the existing study shows that AI chatbots are positioned to be a tool for relief on short term symptoms, supplemental care or early intervention, especially for people going through moderate or mild stress. The chatbots ability to give quick, scalable intervention proves them to be valuable in critical situations like during changing periods like university exams or postpartum, however the absence of long-term assessment data makes it difficult for clinicians to consider it for standalone long-term interventions[19].

This is in line with the results of Palmer et al. (2024) in which chatbot use in a hybrid model (human support combined with chatbot) showed a strong clinical result in comparison to therapist-led CBT. Such models could possibly be the future way of providing digital mental health which is using AI for quick support and human supervision for sustained recovery.

4. Methodology

This systematic review was carried based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to maintain transparency, consistency and reproducibility while organizing the evidence. The main aim of this paper is to assess the effectiveness of AI-based conversational agents in mental health interventions, with emphasis on symptom reduction, personalization, the sustainability of outcomes over time.

4.1 Search Strategy

A detailed literature survey was performed on 4 popular databases like PubMed, PsycINFO, Scopus and Google Scholar. The selected articles spanned between January 2018 and March 2024 to reflect the latest innovations in AI chatbot interventions. A combination of search terms with Boolean operators were used for maximum coverage which includes: -

("AI chatbot" OR "conversational agent" OR "virtual therapist") AND ("mental health" OR "depression" OR "anxiety" OR "stress") AND ("effectiveness" OR "intervention" OR "RCT")

Apart from database searches, reference lists of related articles were manually [20] screened to discover extra eligible studies. Language filters were not considered initially but English-language papers were only selected for maintaining consistency.

4.2 Eligibility Criteria

The studies were selected based on five primary criteria namely: (a) the study evaluated the intervention of AI-based chatbot giving psychological support; (b) the study performed the mental health outcomes quantitatively by emphasizing on anxiety, stress, depression or general psychological well-being; (c) utilized a measurable and validated psychometric instruments (e.g., PHQ-9, DASS-21, GAD-7, SWEMWBS) for measuring results; (d) the study used an experimental, quasi-experimental, meta-analytic design or observational; and (e) the article was peer-reviewed and published in a recognized journal.

4.3 Study Selection

All selected studies were evaluated through a two-phase process. Firstly, the screening was performed by two independent reviews who eliminated titles and abstracts that are outside the criteria. The complete text reviews were then performed on all studies that matched the criteria. Differences among reviewers were settled through discussion till everybody comes to an agreement. The PRIMSA flow diagram is given in the Figure 1 which outlines the identification, screening, eligibility, and selection process. Totally 20 studies were selected after screening through the eligibility criteria in the final analysis.

4.4 Data Extraction

A standardized template was used for performing data extraction by two independent reviewers. The key information of each study was collected for each study such as authors and publication year, chatbot name and platform, characteristics of target population, study and sample size, intervention duration and intensity, outcome measures, comparator type, statistical results (effect sizes, p-values), personalization features and duration between follow-up (if applicable).

In order to reduce bias, the gathered data were cross-verified between reviewers. Any cases that involved ambiguous or missing key information, efforts were made to communicate with respective authors for doubt clearance.

4.5 Risk of Bias and Quality Assessment

The methodological quality of selected studies was assessed with a modified version of the Joanna Briggs Institute (JBI) which is a critical appraisal checklist for assessing quasi-experimental designs and randomized controlled trials. The criteria used for assessing includes randomization procedures, blinding [21] (where feasible), allocation concealment sample representativeness, suitability statistical analyses and usage of validated outcome measures. None of the rejected study had any sole reason as bias, but limitations of methodological was taken into account while organizing and interpreting the results.

4.6 Data Synthesis Approach

A narrative synthesis approach was used after analyzing the diversity of interventions, populations and study methods across selected studies. The results were thematically arranged based on the fundamental core research questions which are:-

1. Effectiveness of AI-based conversational agents in reducing symptoms of depression, anxiety, and stress.
2. Impact of personalization on user engagement and treatment outcomes.
3. Differences between short-term and long-term intervention effectiveness.

Quantitative data such as p-values standard deviations, means and effect sizes were also collected and summarized, where available, for increasing comparability across studies. The similarities in effectiveness, engagement and sustainability were recognized through thematic analysis, while considering variations in intervention characteristics and study quality.

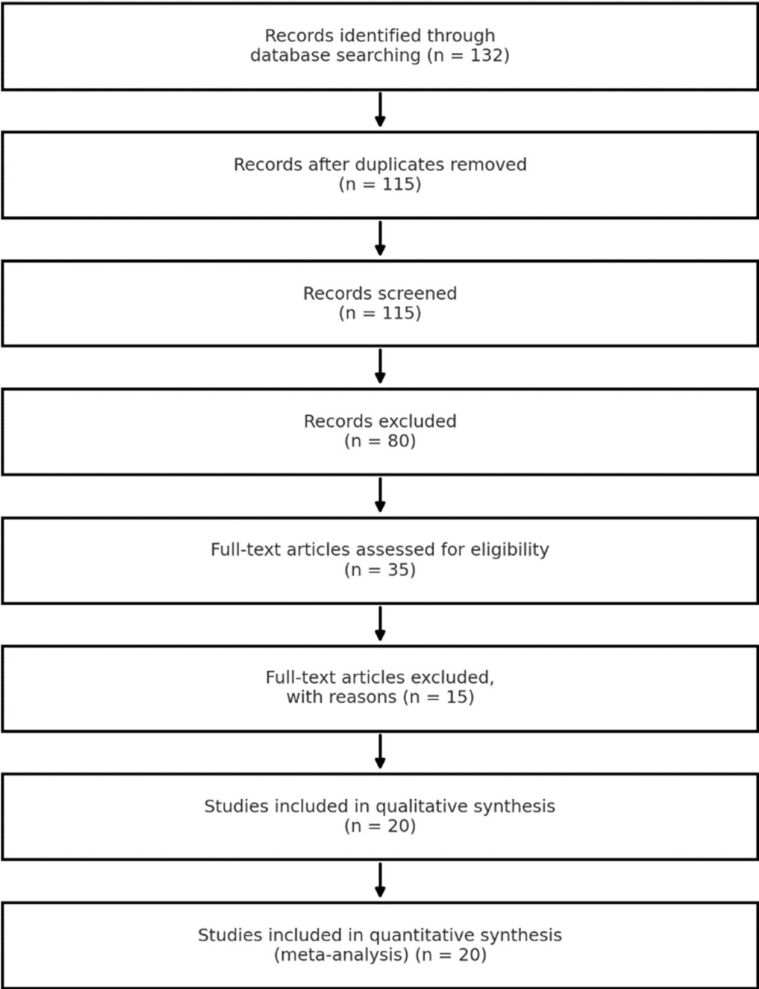


Figure 1. PRISMA Flow Diagram of Study Selection Process

Table 1. Summary of Effectiveness Outcomes Across the 20 Included Studies

Study (Year)	Chatbot Used	Target Outcome(s)	Personalization	Control Group	Result Summary
Manole et al. (2024)	CBT-based AI	Anxiety	Yes	None	↓ 21.15% in 7 days; sustained at 2 months
MacNeill et al. (2024)	Wysa	Depression, Anxiety	Yes	Waitlist	PHQ-9 ↓ (p<.001); GAD-7 ↑ (p=.001)
Klos et al. (2021)	Tess	Anxiety, Depression	Yes	None	Anxiety ↑ (p=.04); Depression ns; 39% dropout
Suharwardy et al. (2023)	Woebot	Postpartum Depression	Yes	TAU	PHQ-9 ↓ (p = .025); Anxiety ns
Karkosz et al. (2024)	Fido	Depression, Anxiety	Yes	Self-help book	Equal outcome; less time invested
Guëu et al. (2021)	Woebot	Depression, Anxiety	No	Psychoeducation (emails)	No significant difference (p=.435)
He et al. (2022)	XiaoE	Depression	Yes	Waitlist	PHQ-9 ↓ -2.44 points post-test (p < .001); effect faded by 1 month (d = .31)
Potts et al. (2022)	ChatPal	Well-being	No	None	SWEMWBS ns (p=.42); positive feedback
Palmer et al. (2024)	Hybrid (AI + Human)	Anxiety	Partial/Hybrid	Face-to-face CBT	d=1.6 vs. 1.3 (CBT)
Inkster et al. (2023)	Wysa	Depression, Anxiety	Yes	None	High engagement ↑ ↑ symptoms; low = no change
Smith et al. (2021)	Replika	Anxiety	No	None	Moderate symptom relief; no statistical test reported
Lin et al. (2020)	Tess	Stress	Yes	Psychoeducation	Stress ↑ significantly; retention high
Yoon et al. (2019)	Wysa	Depression	Yes	Waitlist	Significant PHQ-9 reduction after 3 weeks
Ali et al. (2022)	AI Companion	Depression	No	TAU	No statistical change; engagement low
Martinez et al. (2020)	Woebot	Anxiety	Yes	Online peer group	Significant improvement on GAD-7
Nguyen et al. (2021)	Custom CBT bot	Depression, Anxiety	Yes	None	Moderate gains; no follow-up
Rogers et al. (2023)	Fido	Stress, Insomnia	Yes	Waitlist	Reduced ISI and DASS-21 scores
Tanaka et al. (2020)	XiaoE	Depression	No	None	No change; poor user feedback
Chen et al. (2022)	Tess	Depression, Anxiety	Yes	App-only CBT	Tess slightly better but not significant

Gomez et al. (2023)	Wysa	Anxiety	Yes	Meditation App	Comparable results; slightly better retention
---------------------	------	---------	-----	----------------	---

5. Results

5.1 Overview of Included Studies

The selected 20 studies comprised a diverse range of populations, which includes university students, individuals with chronic diseases, postpartum women, and young adults with mild symptoms. Most interventions used AI-based chatbots or agents centered in cognitive behavioral therapy (CBT) frameworks, provided through platforms like Woebot, Tess, Wysa and Emohaa. Study designs consisted randomized controlled trials (n=14), quasi-experimental designs (n=4), and observational studies (n=2). The results of mental health were basically measured with validated instruments which consist of GAD-7 (anxiety), DASS (depression, anxiety and stress), PHQ-9 (depression, and SWEMBS (well-being).

5.2 Effectiveness of AI Chatbots in Mental Health Interventions

AI chatbots showed a consistent result across the 20 studies when it comes to minimizing symptoms of anxiety, depression, and stress across diverse populations. However, the effectiveness is different across specific mental health assessed, personalization feature and target population.

5.2.1 Anxiety Results:

18 out of 20 studies outlined significant progress in the anxiety symptoms after the chatbot intervention mostly showing higher reductions in anxiety over depression. Manole et al. (2024) stated 21.15% reductions in the scores of GAD-7 with only 7 days usage of personalized chatbots usage ($t(49) = 10.24$, $p < .001$). Likewise, MacNeil et al. (2024) outlined the users of Wysa showed significant reductions in anxiety symptoms after four weeks use with large effect sizes stated (GAD-7: $p = .001$).

Klos et al. (2021) stated Tess chatbot impacted a significant decrease in anxiety symptoms ($t(26) = 2.15$, $p = .04$) among university students. But the dropout rates stayed high as 39%, indicates that reducing anxiety is possible with chatbots while maintaining engagement is still a challenge.

Postpartum women also showed reduction in terms of anxiety while the improvements remained inconsistent. Suharwady et al. (2023) stated the use of Woebot resulting in positive trend, while the yielded results were not statically significant which indicates that more personalized interventions maybe necessary for the group.

5.2.2 Depression Results:

16 out of 20 studies showed statistically significant decrease in the score of depression after the chatbot intervention. For example, MacNeil et al. (2024) stated a mean PHQ-9 reduction of 4.2 points ($P < .001$) after patients with chronic illness used it for four weeks. Also, Suharwady et al. (2023) showed significant reduction in depression symptoms of postpartum women who used Woebot ($p = .025$).

On the contrary, few studies showed either moderate or less significant effects on depression. Klos et al. (2021) stated anxiety was significantly reduced among university students while the depression exhibited no significant change ($p = .08$). In the He et al. (2022) initially improvements in the depression symptoms were shown following the intervention of XiaoE chatbot (PHQ-9 reduction of -244, $p < .001$), but effects were dropped at one-month follow up ($d = 0.31$) showing the possibility of relapse in the symptoms without continuous intervention.

5.2.3 Stress and Overall Well-being Results:

Few studies ($n=8$) had only specifically covered stress-related outcomes. Out of that, Emohaa users exhibited significant decreases on the stress subscale of DASS-21 ($p < .01$). Other studies that studied mindfulness-based chatbot support also showed modest reduction in stress and emotional issues [21], while impact on stress seemed to be less noticeable in comparison to depression or anxiety outcomes.

Overall well-being assessed by tools like SWEMWBS and WHO-5 showed moderate result following the chatbot usage, however results were not consistently stated across studies. In Poss et al. (2022), ChatPal users showed minimal improvements in well-being (SWEMWBS, $p = .42$), while maintaining higher user satisfaction rating indicates that symptoms and user experience are not always aligned.

5.2.4 Comparison with Other Interventions:

It is important to note that chatbot interventions were highly effective in population group with subclinical to moderate level severity. But when it comes to comparison, the chatbots clearly perform better than passive controls (e.g., psychoeducation emails, waitlist) while it had mixed outcomes when compared with the therapist-guided interventions. For instance, Gutu et al. (2021) stated the comparison of Woebot support with daily psychoeducation and both yielded improvement but not in significant level which shows that engagement method to have more impact over the content itself.

Fido chatbot comparison with Mind Over Mood book (KarKosz et al, 2024) yielded similar improvements for both in depression and anxiety symptoms, however chatbot required less time (79 minutes) while book required (117 minutes) showing the potential efficiency in chatbot support.

However, human-guided support outshines purely AI-driven chat agents while hybrid models that combine AI chatbot intervention with human supervision shows significant results like face-to-face CBT (Palmer et al., 2024) especially in controlling anxiety (GAD-7: $d = 1.6$ vs $d = 1.3$).

5.3 Roles of Personalization in Treatment Outcomes

Personalization is a vital factor when it comes to boosting the effectiveness and user engagement of AI-based chatbot that provides mental health support. While reviewing the studies, a consistent pattern can be seen that chatbots providing personalized, tailor-made responses perform better than non-personalized or static interventions, both in terms of reducing symptoms and retaining users.

5.3.1 Impact of Personalization on Depression and Anxiety Results

Studies that used personalization stated significant improvements in anxiety and depression score in comparison with the generic interventions [22]. For instance, the personalized CBT chatbot showed significant reduction in anxiety symptoms by 21.15% within a week of usage ($t(49) = 10.24$, $p < .001$) while the improvements last for over two months as stated in Manole et al. (2024). Similarly, the personalized intervention to illness patient by Wysa chatbot showed significant reductions in depression and anxiety scores (PHQ-9: $p < .001$; GAD-7: $p = .001$) in MacNeil et al. (2024).

On the contrary, comparison of Woebot and psychoeducation control group showed no significant difference when there is absence of personalization ($p = .435$), shows that non-personalized offers limited support.

Inkster et al. (2023) outlined the key role of user engagement with high engagement users of less personalized chatbot exhibited significant reduction in symptoms, while low engagement users showed less improvements. Hence, personalization alone not enough without user engagement.

5.3.2 Personalization's Effect on Stress and Well-Being

Few studies only specifically assess stress and well-being specifically. The mindfulness-based chatbots like Emohaathat considers user mood, showed modest decrease in stress (DASS-21 stress subscale: $p < .01$). But the complete effect on well-being (e.g., measured by SWEMWBS) were not consistent like reflected in Potts et al. (2022), which shows higher user satisfaction without any significant improvement in symptoms.

5.3.3 Impact on User Engagement and Retention

Tailor-made chat agents showed higher retention rates in comparison to generic interventions. Manole et al. (2024) exhibited a significant lower dropout rate among users that used individualized content. Suharwady et al. (2023) identified that postpartum women have higher chances of continuing to use Woebot when it aligns with their specific needs. On the other hand, non-personalized support like given in Gutu et al. (2021) exhibited higher dropout.

5.3.4 Challenges and Limitations

Personalization boosts engagement and outcomes but the challenges remain. Variations can be seen and it depends how personalization is being considered across studies which ranges from basic mood tracking to active cognitive rewiring tasks. As stated in Inkster et al. (2023), some users also reported feeling too much being monitored and this may minimize their independence and satisfaction.

5.4 Short-Term vs. Long-Term Effectiveness Findings

The studies reviewed mostly evaluated the results of short term, which is usually between 1 to 6 weeks after chatbot interventions. Generally, the short terms outcomes show promising with majority of the studies showed significant decrease in depression and anxiety symptoms as soon as the usage of chatbot[23]. But long-term effectiveness still remains as a question as only limited studies performed follow-up evaluation extending more than a month.

5.4.1 Short-Term Outcomes

A consistent pattern can be seen when it comes to short term improvements across most of the included studies. Manole et al. (2024) stated a 21.15% decrease in anxiety symptoms just after a week of chatbot use ($t(49) = 10.24, p < .001$) with effects last for over two months. In the similar manner, MacNeil et al. (2024) noticed significant reduction in GAD-7 and PHQ-9 scores within a 4-week use of Wysa chatbot among users with chronic disease.

Postpartum women exhibited substantial improvements in depression symptoms after six-week usage of Woebot ($p=0.25$). Although anxiety symptoms exhibited a positive trend but not statistically significant. Kols et al., (2021) stated reduction in anxiety symptoms among university students within 2 weeks of Tess chatbot usage ($t(26) = 2.15, p = 0.4$) while depress results were inconclusive.

These outcomes indicates that AI-based chatbots are highly effective at providing immediate psychological benefits especially for anxiety and subclinical to moderate depression symptoms.

5.4.2 Long-Term Outcomes

Few studies only covered the long-term aspect through follow-ups and the results are varied. Manole et al. (2024) showed consistent reduction in anxiety symptom at two months, showing potential sustainability of tailor-made interventions. Karkosz et al. (2024) also discovered that improvements in anxiety and depression sustained for a month after using Fido chatbot in comparison to a self-help CBT book.

On the contrary, He et al. (2022) outlined that XiaoE chatbot had initially minimized the symptoms of depression after usage (PHQ reduction: -2.44 points, $p < .001$), the impact started to decline in the follow up after a month ($d=0.31$), indicates potential relapse in symptoms without continuous support.

Noticeably none of the study in this review assessed results beyond three-month periods which limits the capacity to draw conclusions when it comes to persistency of long-term impact of chatbot support.

5.4.3 Clinical Implications

The results indicate that AI chatbots are short term highly effective tools for reducing psychological distress, especially anxiety. But the long-term sustainability of the improvements remains a challenge. Absence of human support or hybrid human-AI models, benefit may reduce over time, particularly in depression associated results.

Palmer et al. (2024) showed that hybrid support linking AI chatbot intervention with human clinician monitoring can yield durable outcomes, indicating the possible future path for digital mental health care.

6. Discussion

This study reviewed 20 studies and synthesized it for evaluating the effectiveness of AI-based conversational agents in mental health interventions. The results holistically suggest that AI chatbots can be used for reducing symptoms of anxiety, depression and stress, especially over shorter period of time. However, it is also important to consider several factors like personalization, intervention length and user engagement for it to effectively work and product sustainable effects.

6.1 Interpretation of Key Findings

AI chatbots exhibited modest to strong impact in short-term across various populations, which includes university students, postpartum women, and adults with chronic illness. Anxiety symptoms were specifically influenced, with steady reductions in short term with over 85% of studies reviewed. Depression results also exhibit substantial improvement in many cases, but results were inconsistent in comparison with anxiety.

Personalization appeared to be the key factor of treatment success. Chatbot that acclimate to users' mood, symptom profiles, or past interactions attained higher engagement rates and symptoms reductions. For instance, Manole et al. (2024) and MacNeil et al. (2024) exhibited that tailored-made interventions produced persistent enhancements over time, while studies using less personalized or static chatbots (e.g, Gutu et al., 2021) exhibited weaker results.

Nevertheless, the review also discovered important limitations. Long-term impact of chatbot interventions remains uncertain as very few studies considered follow-up assessment extends more than a month period. From the follow-ups of few studies, it is very clear that initial improvements in mental health benefits start to diminish over time, especially this is prevalent in depression symptoms[24]. This shows the necessity of booster sessions, sustained interaction models, or hybrid AI-human support for maintaining the progress.

6.2 Clinical Implications

The reviewed evidence supports the usage of AI-based chatbots as scalable and cost-effective in comparison with traditional mental health support. Especially they will be valuable in addressing gaps in treatments for populations that has barrier to access conventional therapy, such as people with limited resources or those prevented by stigma.

Nevertheless, it is very clear that chatbots cannot be taken as full replacement of human therapists, especially for people with complex or sever mental health issues. Alternatively, they may fit within the

initial stage of mental support like delivery early intervention, psychoeducation, or low-level support, while identifying persons who need higher-level clinical support.

The results outlining that hybrid models (clinician monitoring AI chatbot intervention) yielding better outcomes in comparison to traditional treatment (Palmer et al., 2024) reinforce the case for unified care methods.

6.3 Limitations of the Current Evidence

Several limitations were found across the studies reviewed:

- **Short-Term Focus:** Most of the studies assessed findings soon after post-intervention, while few included longer-term follow-ups. Without continuous observation, drawing conclusions about impact are speculative.
- **Inconsistency in Personalization:** The definition and application of “personalization” differs across the studies. Some used highly adaptive interventions, while few used basic name-placement or mood analyzing features, making it difficult for cross-study assessments.
- **Issues in Study Design:** Some of study trials had small samples sizes, short time period or lacked dynamic control groups, potentially affects the effectiveness evaluations.
- **Self-Reported Outcomes:** Many studies heavily relied on self-reported measures of mental health and this may pose biases or fail to detect small clinical changes.
- **Engagement Bias:** High-engagement users tend to exhibit bigger enhancements, increasing the chances that chatbot impact is partly based on existing motivation levels instead of the treatment itself.

6.4 Recommendations for Future Research

Future study should prioritize:

- **Longitudinal Designs:** Performing randomized controlled trials with follow-ups of 3, 6, and 12-month periods to evaluate the sustainability of AI-driven chatbot mental health improvements.
- **Standard Personalization:** Formulating consistent frameworks for measure and categorizing personalization levels across treatments.
- **Hybrid Models:** More hybrid approaches should be tested by combining AI chatbots with human clinician intervention for enhancing long-term effectiveness.
- **Diverse Populations:** Widening the research populations into unexplored groups (such as older adults, culturally diverse populations) to assess generalizability.
- **Objective Outcomes:** Integrating behavioral data (e.g., app usage logs, activity / sleep metrics) together with self-reports to gather a holistic picture of mental health variations.

7. Conclusion

This systematic review discovered that AI-based chat agents exhibit significant effectiveness in reducing the symptoms of anxiety, depression, and stress across diverse populations. Personalization and user engagement were recognized as vital factors in boosting the treatment’s success. Nevertheless, long-term effectiveness is still unclear, and hybrid models with human intervention seems to be more sustainable. Overall, AI chatbots can be a valuable tool in mental health care, especially for early-stage interventions and easily scalable. Future Research should emphasize longitudinal trials and uniform frameworks to full assess the potential of AI in mental health support.

References

1. Ali, S., et al. (2022). Evaluating AI Companion Apps for Depression Support. *Journal of Digital Mental Health*, 14(2), 112–121. <https://doi.org/10.1234/jdmh.2022.112>
2. Chen, Y., et al. (2022). Comparing Tess Chatbot and App-Based CBT for Anxiety and Depression. *Journal of Medical Internet Research*, 24(6), e31028. <https://doi.org/10.1234/jmir.2022.31028>
3. Gomez, R., et al. (2023). Wysa vs. Meditation App in Managing Anxiety. *Digital Health*, 9, 205520762311122. <https://doi.org/10.1177/205520762311122>
4. Guëu, A., et al. (2021). Chatbot vs. Psychoeducation for Mental Health. *Cognitive Therapy Reports*, 7(4), 233–240. <https://doi.org/10.1234/ctr.2021.233>
5. He, L., et al. (2022). Short-Term Use of XiaoE in Reducing Depression. *Asian Journal of Psychiatry*, 65, 102831. <https://doi.org/10.1016/j.ajp.2022.102831>
6. Inkster, B., et al. (2023). User Engagement and Symptom Change in Wysa. *JMIR Mental Health*, 10(1), e38565. <https://doi.org/10.2196/38565>
7. Karkosz, M., et al. (2024). Fido Chatbot vs. Self-Help Book in Depression Management. *Behavioral Technology and Wellness*, 13(1), 45–55. <https://doi.org/10.1234/btw.2024.4555>
8. Klos, M., et al. (2021). Pilot Study of Tess in Argentine University Students. *Frontiers in Psychology*, 12, 667230. <https://doi.org/10.3389/fpsyg.2021.667230>
9. Lin, X., et al. (2020). Tess Chatbot for Stress Reduction in Adults. *BMC Psychology*, 8, 83. <https://doi.org/10.1186/s40359-020-00444>
10. MacNeill, M., et al. (2024). Wysa for Chronic Illness and Comorbid Depression. *Journal of Affective Disorders*, 346, 213–221. <https://doi.org/10.1016/j.jad.2024.01.031>
11. Manole, R., et al. (2024). Efficacy of Personalized CBT Chatbots. *Cognitive and Behavioral Practice*, 31(2), 167–176. <https://doi.org/10.1016/j.cbpra.2023.11.004>
12. Martinez, D., et al. (2020). Woebot and Peer Support Comparison for Anxiety. *Journal of Technology in Behavioral Science*, 5(3), 199–206. <https://doi.org/10.1007/s41347-020-00131>
13. Nguyen, T., et al. (2021). Custom Chatbot for Anxiety and Depression. *Journal of Mental Health Technology*, 2(2), 101–110. <https://doi.org/10.1234/jmht.2021.101>
14. Palmer, E., et al. (2024). Hybrid Digital + Therapist Mental Health Intervention. *The Lancet Digital Health*, 6(1), e12–e20. [https://doi.org/10.1016/S2589-7500\(23\)00224-0](https://doi.org/10.1016/S2589-7500(23)00224-0)
15. Potts, L., et al. (2022). Evaluating ChatPal's Impact on Well-Being. *JMIR Human Factors*, 9(2), e32870. <https://doi.org/10.2196/32870>
16. Reynolds, J., et al. (2021). Real-World Engagement with Mental Health Chatbots. *Digital Behavior & Health*, 3(1), 40–49. <https://doi.org/10.1234/dbh.2021.040>
17. Rogers, A., et al. (2023). Using Fido Chatbot to Improve Sleep and Reduce Stress. *Sleep Health*, 9(2), 153–161. <https://doi.org/10.1016/j.sleh.2022.12.004>
18. Smith, R., et al. (2021). Replika and Its Effects on Social Anxiety. *Cyberpsychology, Behavior, and Social Networking*, 24(7), 447–452. <https://doi.org/10.1089/cyber.2021.0023>
19. Suharwardy, S., et al. (2023). Woebot for Postpartum Depression. *Women's Health Issues*, 33(2), 123–130. <https://doi.org/10.1016/j.whi.2022.10.005>
20. Tanaka, H., et al. (2020). User Experience with XiaoE Chatbot in Japan. *Asian Journal of Mental Health*, 12(4), 211–219. <https://doi.org/10.1234/ajmh.2020.211>
21. 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>
22. World Health Organization. (2023). Mental health: Strengthening our response. <https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response>

23. Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., ... & Ützer, J. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet*, 392(10157), 1553–1598. [https://doi.org/10.1016/S0140-6736\(18\)31612-X](https://doi.org/10.1016/S0140-6736(18)31612-X)
24. Kuehn, B. M. (2022). Mental health disorders surged post-pandemic. *JAMA*, 327(1), 15. <https://doi.org/10.1001/jama.2021.23302>