OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Benefits and Risks of AI in Psychotherapy

Shruti Shyam
shruti.shyam09@gmail.com

ABSTRACT

Objectives: This systematic review examines the feasibility and safety of healthcare chatbots while also addressing the ethical implications of integrating artificial intelligence (AI) into clinical care.
Methods: A PubMed search was conducted utilizing keywords such as "Artificial Intelligence (AI)," "psychotherapy," and "chatbot." The papers were then processed through two rounds of screening and evaluated for credibility by the Joanna Briggs Institute (JBI) critical appraisal checklist.
Results: Results of controlled experiments regarding Artificial Intelligence's implementation in psychotherapy sessions in various settings ultimately state that chatbots have the ability to provide varying levels of relief from common mental health issues.
Discussion: Limitations of implementing AI in psychotherapy include a lack of engagement, depth, and emotional regulation, a high risk of inaccuracy, and various data concerns. Gradual integration of AI would be most beneficial, as well as the involvement of a human therapist to ensure patient safety and assess progress.

INTRODUCTION

Healthcare processes are constantly being evaluated, and new practices are often implemented to maintain pace with the population, improve ease and efficiency, and extend healthcare to marginalized groups on a global scale. Artificial intelligence at the surface level is greatly beneficial, as it has the capacity to broaden access to mental health treatment geographically and economically, provide instant feedback during an emergency, allow endless customization for both the user and therapist, and reference an infinite amount of resources to treat patients in a fraction of a second. Online-based or AI-based supports have the potential to eliminate stigma-associated barriers to mental health care and help people seek their desired form of assistance.

Since artificial intelligence chatbots have been available to the general public in many forms, including role-playing games (Sobowale, K., and D. K. Humphrey) and social media, often accessed by the younger generation (Hudon, A. et al.), documented cases about AI's unsafety and inconsistencies pose questions about whether these models should be trusted to aid serious mental health issues. Additionally, recent technological advances call for a more in-depth analysis of new technology's influences on traditional,

refined practices. More specifically, the rise in artificial intelligence and its potential to revolutionize healthcare.

While many case studies have tested developing chatbots, few have compared the chatbots' accuracy, scalability, and ability to replicate human connection. The purpose of this study was to examine these healthcare bots' feasibility, safety, and address ethical concerns with AI chatbots in psychotherapy. This study is in the form of a systematic review, as it has the ability to synthesize relevant data to answer current questions and introduce new inquiries to better refine healthcare technologies.

METHODS

A systematic review was conducted to address the following inquiries about AI in psychotherapy:

- What are possible therapeutic models implemented by AI chatbots?
- How affordable and scalable are AI-supported psychotherapy tools compared to human-delivered therapy?
- Can AI systems demonstrate empathy in a way that is perceived as meaningful by clients?
- Who is responsible when AI systems cause harm or provide incorrect guidance?

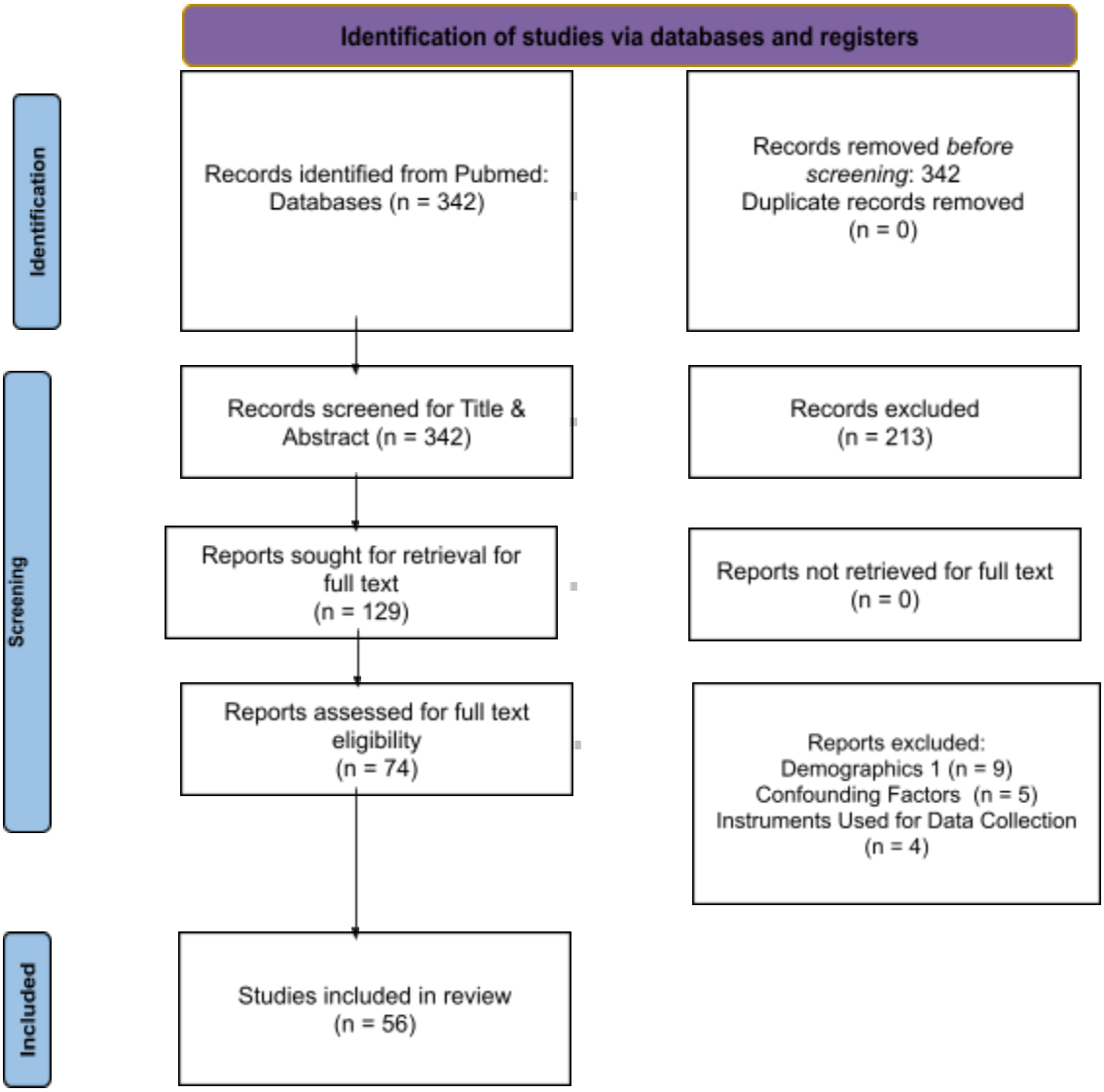
A PubMed search was conducted, filtered based on concepts such as Artificial Intelligence, psychotherapy, talk therapy, mental health intervention, and papers were not excluded based on neither population nor geographical context. 342 papers published between the years 2023 and 2026 were found using the following keywords:

(("Artificial Intelligence"[MeSH Terms] OR "Artificial Intelligence"[Title/Abstract] OR "AI"[Title/Abstract] OR "machine intelligence"[Title/Abstract] OR "intelligent systems"[Title/Abstract] OR (("chatbot*"[Title/Abstract] OR "chat bot*"[Title/Abstract] OR "chat bot*"[Title/Abstract] OR "conversational agent*"[Title/Abstract] OR "conversational system*"[Title/Abstract] OR "virtual agent*"[Title/Abstract] OR "dialogue system*"[Title/Abstract] OR "conversational AI"[Title/Abstract] OR "therapeutic chatbot*"[Title/Abstract]) AND ("loattrfree full text"[Filter] AND "loattrfull text"[Filter] AND "humans"[MeSH Terms] AND "english"[Language]))) AND ("Psychotherapy"[MeSH Terms] OR "Psychotherapy"[Title/Abstract] OR "psychological therapy"[Title/Abstract] OR "talk therapy"[Title/Abstract] OR "mental health intervention*"[Title/Abstract])) AND ((ffrft[Filter]) AND (fft[Filter]) AND (humans[Filter]) AND (english[Filter]) AND (2023:2026[pdat])) on December 22, 2025.

After screening for Title & Abstract, 129 papers remained and after the second, full text screening, 74 papers remained. Papers were screened in the first round based solely on their abstracts, whether or not they related to the use of AI in psychotherapy, or any mental health related procedure. An emphasis on variety was given in order to better evaluate to what extent AI can and should be utilized regarding mental well being. Finally, 18 were subsequently eliminated using the JBI appraisal checklist, disqualified due to

demographics unfit for this paper, possible confounding factors being present, and unsuitable methods of data collection. A total of 56 papers were used for this synthesis.

RESULTS



From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

Figure 1: Flowchart outline of the database search and screening process.

About 22 of 56 synthesized papers are Randomized Controlled Trials (RCT), as they are well-suited for evaluating efficacy and comparative outcomes of chatbot assisted psychotherapy. For example, in one study (Six, S. et al.), the experiment consisted of 4 conditions based on conversational or animation features (conversational and animated, conversational and non-animated, animated and conversational, and non-conversational and non-animated). The nature of this experiment allowed for the use of mixed effects, such as ANOVA, to analyze depressive, stress, and rumination symptoms in each of the four groups.

About 10% of the papers are cross-sectional studies or surveys, meant to reflect a user's overall experience with the chatbot. In one study (Sobowale, K. et al.), trained raters prompted 5 Gen-AI chatbots mimicking the persona of a youth with mental health challenges. They evaluated these chatbots using the Conversational Agent for Psychotherapy Evaluation II (CAPE-II) framework; the primary outcomes were rated in 9 sections, and each section was compared between chatbots using Bonferroni-corrected chi-square tests.

Another 10% include systematic reviews, useful for a big picture understanding of multiple studies. One study (Bucher, A. et al.) synthesizes a variety of Large Language Models (LLM) conversation agents across multiple settings and interfaces and their usefulness in mental health care.

The remaining 20 papers or so are mixed methods, feasibility, or pilot studies, which are useful for early-stage evaluation and recognizing potential benefits of chatbots, and finally, qualitative papers, which address ethical concerns, the science behind chatbot therapy, and improving therapy chatbot designs.

Population/Sex

From the total 56 academic papers used, 48 relied on data collection through participants, and 8 were scoping reviews and integrative reviews. *A Prompt Engineering Framework for Large Language Model-Based Mental Health Chatbots: Conceptual Framework* introduces a conceptual model for prompt engineering and proposes a design that integrates evidence-based therapeutic models, adaptive technology, and ethical safeguards. One study (Bucher, A. et al.) introduces a 3-layer morphological framework to structure and analyze how LLMs are applied, with the goal of informing future research and design for more effective mental health interventions.

29 of the population studies consist of adults (18+), 17 studied adults with conditions such as mild cognitive impairment, 6 studied adults in more severe mental states, and 6 specified no mental health conditions. One study that focused on adults with severe mental states (Reyes-Portillo, J. A. et al.) was a randomized control trial conducted at a university, and participants were required to have elevated depression (Patient Health Questionnaire-2 [PHQ-2] score ≥3) or anxiety symptoms (Generalized Anxiety Disorder-2 [GAD-2]; score ≥3).

Conversely, in (Held, P. et al.), the participating individuals were recruited via social media and word of mouth, and were eligible as long as they were at least 18 years of age, could read and write English at least at a sixth-grade level, and had access to an internet-capable device.

8 of the population studies examine specific subpopulations. One of such papers (Fietta, V. et al.) was a protocol for a pilot feasibility study to examine the effects of a structured chatbot-based dialogue (ALBA) on women dealing with mental health adjustments as they undergo breast cancer treatment. Another paper concentrated on full-time employees of Sony Group Corporation in Japan who demonstrated levels of subthreshold depression.

Setting

About 32 of 56 of studies are conducted in community-based, remote, or online settings, with participants typically accessing chatbot interventions from home via web or mobile platforms. 30 of these 32 studies explicitly describe fully remote designs, such as online digital interventions in the United States or community-based chatbot use in countries like South Africa and New Zealand, highlighting the scalability and accessibility of these tools outside traditional clinical environments.

A substantial portion (~80%) of the dataset also includes mixed or multi-setting studies, combining clinical, community, and remote contexts across different countries. For example, 30 of these reviews and multi-study analyses synthesize data from outpatient clinics, community samples, and digital platforms across regions such as Europe, Asia, and North America, reflecting the global and heterogeneous nature of AI mental health research.

A smaller but notable group (~ 15%) of studies takes place in clinical settings, particularly outpatient environments. Examples include a psychosomatic outpatient clinic in Taiwan, a breast cancer outpatient clinic in Italy, and an outpatient pediatric setting at the Mayo Clinic in the United States. These settings often involve more structured interventions and closer alignment with traditional healthcare delivery.

About 10% of the studies also occur in university or experimental research settings, often involving non-clinical populations such as students. For instance, some U.S.-based lab studies and university-based trials in China examine chatbot use in controlled or semi-controlled environments, blending real-world interaction with experimental oversight.

Additionally, a moderate portion (26 out of 56) of studies spans diverse international contexts, including countries such as the United States, China, Italy, the United Kingdom, Japan, Jordan, Ukraine, and South Africa. Some studies (3 out of 56) even take place in unique or high-need environments, such as outpatient care in conflict zones like Ukraine, emphasizing the potential reach of AI tools in underserved or crisis-affected populations.

Finally, a limited number of papers (8 out of 56) are non-empirical (e.g., reviews or conceptual discussions) and therefore do not specify a single setting, instead discussing applications across clinical and community environments. Overall, the dataset shows a strong trend toward remote, community-based deployment, with relatively fewer studies grounded in traditional in-person clinical settings.

Sample Size & Attrition

About half (27 out of 56) of the studies include moderate sample sizes, typically ranging from around 100 to 250 participants in randomized trials, with several examples such as N=210, N=206, and N=223. A substantial portion also includes large-scale datasets or aggregated samples, such as studies analyzing 1,487 participants, 1,901 pooled participants across 15 RCTs, or even 45,876 treatment records, highlighting the growing use of big data in this field.

A smaller group (15 out of 56) consists of small pilot or feasibility studies, often with fewer than 50 participants. For example, some studies included 23 participants, 34 enrolled (with 26 completing), or 50 children, reflecting early-stage testing and prototype evaluation.

In terms of attrition, moderate dropout rates are common, reporting attrition rates of around 18–23%. For instance, one study reported 26% attrition among 50 children, while others showed ~20–23% dropout in samples of over 200 participants. A substantial portion of studies, however, do not report attrition clearly, limiting transparency and making it difficult to assess retention and engagement.

Several studies (close to 35) report minimal or no attrition, particularly in qualitative, survey-based, or short-term designs where dropout is less applicable. Examples include interview studies with 15 participants or surveys with no longitudinal follow-up.

Finally, a notable portion (~45%) of the dataset consists of reviews, protocols, or conceptual papers, where sample size and attrition are not applicable or vary across included studies. Overall, while many interventions achieve moderate sample sizes, inconsistent reporting and relatively high attrition rates remain key limitations in evaluating the reliability and scalability of AI-based mental health tools.

Deployment Context

The majority of studies (39 out of 56) are conducted in a research context, most commonly randomized controlled trials or experimental evaluations of AI mental health chatbots. A substantial portion also includes reviews and conceptual papers, which synthesize existing evidence or discuss potential applications without deploying a new intervention (11 out of 56), 9 of them examining policy, regulation, or implementation contexts, such as discussions on governing AI use in psychotherapy or integrating these tools into clinical care. A smaller but notable group (5 out of 56) involves real-world or exploratory deployments, such as observational studies of existing chatbot use or platforms supporting patients between therapy sessions.

AI Architecture and Intelligence Level

32 of the reviewed papers used LLM or Natural Language Processing(NLP)-based conversational agents. About 17 of these papers relied on rule-based or hybrid systems, and a smaller subset of 7 incorporated advanced machine learning models (Supporter Vector Machine (SVM), Random Forest) or multimodal approaches. Emerging innovations— such as multi-agent architectures, VR-integrated coaching (eg, "Otti" (Hong, J. et al.)), and socially assistive robots— appear in a minority of studies but reflect growing experimentation with autonomous or immersive therapy delivery. Commercial platforms like Replika, Character.AI, and Woebot demonstrate real-world tools. Additionally, hybrid clinical models such as Parent-Child Interaction Therapy (PCIT) integrated models augmented with real-time AI alerts highlight a clear trend towards AI-assisted rather than fully autonomous care.

The papers show a preference towards generative AI, evident as they allow a greater amount of personalization and emotional support. This type of virtual care is unique because it has the ability to replicate a real therapy conversation without a therapist being present. However, complete reliance on Generative AI raises ethical concerns regarding misinformation and emotional dependence. A safer route would include a hybrid therapist-chatbot treatment as intended by the rule-based system. This form of care blends the flexibility of an AI tool while allowing a trained therapist to customize and monitor the patient's conversations and progress. The multimedia tools are less popular as they require a greater amount of resources. Overall, the papers indicate that AI-assisted psychotherapy could be a beneficial supplement to long-term mental health care.

Conversational Design/Interface

Roughly 70% of the papers in this dataset are chatbot-based, with over half featuring open-ended or semi-structured conversational agents delivering Cognitive Behavioral Therapy (CBT)-informed therapeutic support through natural language dialogue. For example, (Reyes-Portillo, J. A. et al.) evaluated the feasibility of a chatbot called Wayhaven. Wayhaven is a wellness chatbot that tailors its content to each individual based on the user's real-time needs. It utilizes the user's demographics, university resources, and ensures safety through advanced crisis detection and rerouting.

A substantial portion (about 30%) also uses structured or semi-structured therapeutic designs (including CBT or guided coaching frameworks), representing a significant subset that emphasizes guided interaction rather than fully open conversation. For example, ALBA is a chatbot-delivered stress management app with 5 preset coaching sessions tailored for women with breast cancer. The app is meant to demonstrate high accessibility as well as engagement, aligning with the World Health Organization Self-Help Plus protocol.

A smaller but notable group (6 out of 56) consists of non-conversational or hybrid systems (e.g., physiologic monitoring tools, VR meditation, robotics, or interactive media generation), indicating that

only a minority of studies move beyond chatbot-style interfaces. In the case of [Dallison S, Moffett B, et al[6]], artificial intelligence is used solely to generate content to personalize the user's experience. Participants were guided to use one of 3 versions of GenAI tools: ChatGPT, MidJourney, and Soundful. ChatGPT was meant to generate stories, MidJourney relied on text-to-image software, and Soundful created music based on user input.

Additionally, several papers (24 out of 56) are conceptual discussions or literature reviews of conversational AI and digital mental health tools, making up a moderate portion of the dataset without presenting primary interventions. (Andersson, G.) discusses the emergence of artificial intelligence tools and its use for advancing digital CBT platforms. It puts a larger emphasis on current iCBT practices, particularly noting the importance of effectively determining 'what works for whom.'

Finally, only a limited number (6 total) of studies focus on surveys, simulations, or analytics models without delivering an actual therapeutic chatbot, representing a small fraction of the overall data. (Boit, S., and R. Patil.) provides a conceptual framework for real-world LLM chatbot development, utilizing a layered architecture combining evidence-based therapeutic models, safety filters, a personalized user state database, and retrieval-augmented guidance. (Stephenson, C. et al.) surveys people's preference/engagement with chatbots by means of epistemic trust, personality, and fear of intimacy.

Therapist Involvement/Interface

The majority of studies (20/56) involve no therapist participation, with most interventions delivered through fully autonomous AI chatbots. (Sobowale, K., and D. K. Humphrey) identifies 4 popular chatbots from OpenAI's GPT store to conduct this study. Two reviewers independently applied the CAPE (Conversational Agent for Psychotherapy Evaluation) framework to these chatbots, using 2 fictional personas to simulate interactions.

A substantial portion (15 out of 56) still reflects variation in therapist involvement, with some studies including optional clinician oversight or mixed human-AI support. (de Graaff, A. M. et al.) utilized a combination of a chatbot conversational interface and guidance from an online, trained e-helper. This support was given to ensure progress was being made and to educate the participants (young adults aged 18-21) on psychoeducational practices.

A smaller but notable group (11 out of 56) incorporates human-in-the-loop or blended care models, where therapists monitor progress or assist during sessions. Participants in (Prasad, N. et al.) interacted with a web-based, internet-delivered CBT platform. To gauge improvements that needed to be made to the model, trained therapists monitored any repeated symptoms during the treatment.

Only a limited number of studies (~ 9%) feature fully clinician-led interventions, typically using AI as a supplementary tool rather than the main treatment.

The 244 patients receiving group-based CBT in 5 of the United Kingdom's National Health Service Talking Therapies in study (Habicht, J. et al.) were encouraged to utilize the AI materials between clinical sessions.

Finally, a small fraction (~3%) consists of conceptual discussions or reviews that describe potential therapist involvement without implementing it in practice. (Jurblum, M., and R. Selzer.) provides a conceptual discussion of AI in psychotherapy and potential models of autonomous chatbots, arguing it has the ability to "provide engaging clinical interactions given its capacity for highly realistic interaction as well as its high-level cognitive and emotional capabilities". In contrast, (Rubin, M. et al.) emphasizes the importance of human empathy in a field such as psychotherapy, something AI significantly lacks. Although studies have demonstrated AI has the ability to generate empathic responses, the author argues that the treatments some conditions require "[transcend] the mere reception of an ideal empathic response" and require conscious human care

Duration of Intervention & Session Frequency / Dosage

The majority of interventions (21 out of 56) last between 4 and 8 weeks, making multi-week programs the most common duration in the dataset. A substantial portion also includes flexible or unspecified durations, particularly in studies allowing naturalistic or participant-driven use rather than fixed timelines. A smaller but notable group involves shorter-term interventions, such as 1-week or brief experimental sessions. Longer interventions (e.g., 2–6 months) appear less frequently, representing a minority of studies. In terms of session frequency, most studies rely on unstructured or participant-driven engagement, with users interacting freely rather than following a strict schedule. A substantial portion still includes light guidance, such as recommended weekly check-ins or daily short interactions (e.g., 5–10 minutes). A smaller group (11 out of 56) uses more structured dosing, such as weekly therapy sessions or formal CBT modules. Only a limited number of studies implement highly intensive schedules, such as daily use combined with multiple weekly therapy sessions.

Symptom Outcomes

The majority of studies (30/56) assess depression and anxiety symptoms, making these the most common outcome measures across the dataset. A substantial portion also examines broader mental health outcomes, such as psychological distress, well-being, stress, and overall quality of life. A user-based Gen-AI chatbot, engaged with regularly, used to target depression and anxiety symptoms, showed greater knowledge in therapeutic treatment as compared to human participants, however lacking in a category known as "radical genuineness." Participants of study (Thunström, A. O. et al.) remarked on the lack of nuance their chatbot was capable of, despite endless customizations they were allowed to make. This indicates that although the chatbot is capable of processing information regarding symptoms, its execution falls short of the connection and trust needed to deliver proper psychotherapy treatment.

A smaller but notable group (15/56) includes more specific or targeted outcomes, such as sleep quality, body image, emotional regulation, or physiological measures. These studies were included to further discuss the capability of the chatbots to deliver meaningful conversations about delicate topics. The results of study (Zhang, X. et al.) focused on body image awareness for young adults. The participants found better success without the AI agent, noting the model faced drawbacks with facilitating body image awareness and self-reflection, availability of conversational support, and discontinuities in user engagement. Finally, several studies (23/56) report non-clinical outcomes, such as user perceptions, therapeutic alliance, engagement, or attitudes toward AI, and a limited number of studies include no direct symptom outcomes.

Effect Metrics

The majority of studies (37/56) report quantitative effect metrics, most commonly moderate effect sizes (e.g., Cohen's d ≈ 0.4–0.5), showing reductions in depressive symptoms. A substantial portion (21/56) also includes statistical significance testing, such as t-tests, Analysis of Variance (ANOVA), and p-values, indicating meaningful pre–post or between-group differences. A smaller but notable group reports advanced or alternative metrics, including odds ratios, mediation effects, risk ratios, and machine learning performance measures like accuracy or Area Under the Receiver Operating Characteristic (AUROC). Several studies present meta-analytic or synthesized effect sizes, forming a moderate portion of the dataset. Finally, a limited number of studies report no formal effect metrics, relying instead on qualitative findings, narrative descriptions, or conceptual discussions.

Durability

Most studies (33 out of 56) in this dataset report improvements in depression and anxiety outcomes, typically measured using standardized clinical scales and supported by moderate effect sizes. These interventions are largely delivered through fully autonomous, CBT-based chatbot systems with flexible, user-driven engagement over short-term periods, most commonly lasting 4–8 weeks. While results consistently show reductions in psychological distress and improvements in well-being, the majority of studies rely on post-intervention assessments without long-term follow-up. This lack of durability data represents a key limitation, as it remains unclear whether the observed benefits are sustained over time or diminish after the intervention ends. Additionally, variation in study design, outcome measures, and therapist involvement further complicates direct comparisons across findings, highlighting the need for more standardized and longitudinal research in this field.

User Experience

The majority of studies (~80%) report generally positive user experiences, with high satisfaction, perceived usefulness, and acceptability of AI mental health tools. For example, users of AI music and image generation tools reported strong engagement, with 91% saying music creation improved their mood and 100% preferring AI-generated images over app-based alternatives. A substantial portion of studies

also highlights qualitative feedback showing users value accessibility, autonomy, and nonjudgmental interaction; for instance, participants described Socrates 2.0 as "enjoyable and engaging," noting that its on-demand support encouraged self-reflection. However, users also reported limitations such as repetitive questioning and limited conversational depth, which sometimes reduced engagement. A smaller but notable group (2#) uses formal usability scales like SUS (System Usability Scales) and UTAUT (Unified Theory of Acceptance and Use of Technology), generally finding moderate to high usability and helpfulness, while also identifying issues like technical glitches or login barriers. Finally, a limited number of studies (~15%) report dropout reasons, most commonly loss of interest or time constraints, rather than dissatisfaction with the chatbot itself.

Technical Limitations or Failures

The majority of studies (34/56) identify AI and chatbot limitations related to conversational quality, especially issues like repetitive, generic, or irrelevant responses. For example, adolescents using Socrates 2.0 reported a need for more content depth and noted its "repeated questioning style and limited conversation depth," which reduced engagement.

A substantial portion of studies (29/56) also highlights technical and design limitations, such as limited contextual understanding, weak memory across conversations, and inconsistent response quality. Several chatbots were described as producing occasional off-topic or inaccurate replies, reflecting early-stage language model constraints.

Methodological limitations in the studies themselves, including small or homogeneous samples, lack of follow-up, cross-sectional designs, and heterogeneity of AI tools, are also discussed. For example, some studies used single-clinic samples or lacked control-group engagement tracking, limiting generalizability and causal inference.

Several papers (19/56) also emphasize system and implementation barriers, such as login issues, connectivity problems, multi-device integration challenges, and prototype-stage instability. These issues affected usability and consistent access to the interventions in real-world settings.

Finally, a limited number of studies (6/56) discuss broader conceptual or ethical limitations, including concerns about hallucinations, overestimating AI empathy, a lack of understanding compared to human therapists, and risks around safety, accuracy, and trust in AI-driven mental health care.

Ethical Concerns / Safety

The majority of studies identify privacy, transparency, and responsible data use as the central ethical concerns in AI-based mental health interventions. Many papers emphasize risks related to how user data are collected, stored, and used, as well as a lack of transparency around model training and decision-making processes. Concerns about data governance are especially prominent in systems using

large clinical or behavioral datasets, with several studies warning that unclear data practices could undermine user trust and safety.

A substantial portion of the literature also highlights clinical safety risks and overreliance on AI, particularly the potential for chatbots to provide inaccurate, inappropriate, or insufficient responses in sensitive situations. Multiple studies raise concerns about improper crisis handling, hallucinations, and the inability of AI systems to manage complex mental health conditions. In response, some interventions explicitly remind users of the nonclinical nature of the tool and provide guidance for seeking professional help. Additionally, several papers note the risk of emotional dependence on AI, where users may begin to rely on chatbots for support in place of human care, raising concerns about long-term psychological impacts.

Another major theme involves bias, fairness, and inclusivity, with studies pointing out that AI systems may reflect or amplify existing biases in training data. This includes concerns about cultural insensitivity, lack of representation, and reduced effectiveness for diverse or underserved populations. Some papers stress the importance of co-design with users and the need for culturally adaptive systems to ensure equitable access and outcomes.

A moderate portion of studies discuss ethical safeguards and research governance, including institutional review board (IRB) approval, informed consent, and adherence to established ethical guidelines such as the Declaration of Helsinki. For example, some studies detail obtaining written parental consent and child assent, while others describe continuous monitoring of chatbot interactions using safety modules. These measures demonstrate attempts to align AI interventions with traditional clinical research standards, although they are not uniformly applied across all studies.

Several papers also emphasize accountability, regulation, and the need for oversight, arguing that AI systems in mental health require clear frameworks to define responsibility for errors, harm, or misinformation. This includes calls for clinician involvement, human-in-the-loop systems, and regulatory standards to ensure safe deployment. Concerns about the impact of AI on clinical practice are also noted, such as the potential erosion of therapist skills or ambiguity in responsibility when AI systems are used in care delivery.

Finally, a smaller subset of studies addresses broader conceptual and ethical challenges, including trust in AI-generated empathy, the limits of machine understanding of human emotion, and the ethical implications of simulating therapeutic relationships. These discussions often highlight the tension between scalability and ethical responsibility, emphasizing that while AI can expand access to care, it should augment rather than replace human clinicians. Overall, the literature consistently underscores that ethical considerations are paramount to the development, evaluation, and implementation of AI in mental health. Steps have already been taken to regulate the use of Artificial Intelligence in psychotherapy on a state level, as in Illinois

Public Act 104‑0054 addresses its implications in administrative support, supplementary support, and therapeutic communication.

DISCUSSION

Healthcare systems are continually reassessed, with new practices introduced to keep up with population needs, improve efficiency, and expand access to underserved communities worldwide. However, recent technological advances require closer examination of how emerging tools affect established clinical practices, particularly the growing role of artificial intelligence in healthcare transformation. On the surface, AI offers significant benefits, including expanded geographic and economic access to mental health care, rapid response in emergencies, highly personalized support for users and clinicians, and near-instant access to vast informational resources. AI and online-based interventions may also reduce stigma-related barriers, encouraging individuals to seek care in ways that suit them. This study examines the feasibility and safety of healthcare chatbots while also addressing the ethical implications of integrating AI into clinical care. Results of controlled experiments regarding Artificial Intelligence's implementation in psychotherapy sessions in various settings suggest that chatbots have the ability to provide relief from common mental health issues. This statement is generally valid throughout the papers, regardless of therapist involvement during the trial, the medium of AI utilized (chatbots, LLMs, iCBT), population, setting, and duration. However, the magnitude of relief given varies throughout the papers; some papers note "significant" increases in overall wellness, while other papers only mention small indications of improved well-being. Although the most common type of illness being treated was depression or anxiety, the trials indicated that Artificial intelligence could treat symptoms of psychological distress, stress, emotional regulation, insomnia, and even physiological measures. The focus on mental health broadened slightly during the research process, and other forms of neurological and physiological treatments were considered.

Other reviews on this topic have proposed models that address ethical and durability risks. (Poudel, U. et al.) acknowledges the efficiency and accessibility of AI chatbots, but perceives a risk when it comes to ethics, bias, and reduced human interaction. Therefore, the author encourages blended AI and human care. The author describes the recent implementations of artificial intelligence as a "paradigm shift in mental health care," but emphasizes the role of human ethics and dignity as the driving force and restraint for the future development of AI in psychiatry. (Tabassum, A. et al.) claims that implementing artificial intelligence has been proven to improve the quality of life for psychotherapy patients, and provides vital support supplementary to professional assistance. However, it mentions the spread of misinformation as a result of the synthesized data that the system functions on, which is extremely harmful in a psychotherapy setting. The paper cautions against trusting LLMs fully, and suggests implementing a "nuanced approach that considers human involvement, contextual factors, and the long-term implications of various design choices." (Tabassum, A. et al.) comprehensively evaluates AI and XR applications and focuses on therapeutic potential, ethical concerns, and societal impacts. The deployment of AI-XR applications improves patient engagement, especially in resource-limited areas. However, the author suggests studies must evaluate their potential across diverse mental health conditions and populations, and raises questions

about data ownership, bias, and privacy risks. (Wang, L. et al.) provides a description of the capabilities of GenAI in mental health applications, especially in areas of psychoeducation. Although the review notes significant gaps exist in these models, particularly when it comes to cultural competency and diagnosis. The author proposes that future research should adopt more diverse evaluation techniques to better address AI's potential in mental health care. In summary, to maximize the benefits and reduce the risks of the common forms of AI in terms of intelligence and model, human therapists should also be monitoring the interactions to ensure safety for the patient.

AI models have the potential to expand access to mental health care across geographic and financial barriers, deliver immediate support in urgent situations, offer extensive customization for both users and clinicians, and draw on vast resources to inform care within seconds. Digital and AI-based tools can also reduce stigma-related obstacles, making it easier for individuals to seek the kind of support they prefer. Evidence indicates that, when implemented with appropriate guidance, these approaches are associated with reductions in depression and anxiety symptoms. Of the prototypes of AI psychotherapy models, users highlighted technical and design limitations they observed. The models were described as lacking depth and engagement, and on numerous occasions, chatbots produced off-topic or inaccurate replies. Future models must monitor the chatbot's interpretation of the information it is presented to terminate the inconsistencies present. Relying completely on artificial intelligence brings up concerns about patient safety. Safety concerns include misinformation, accuracy, lack of empathy, and possible deeper, psychological problems that may pose a struggle in the long term.

Regarding ethical concerns and privacy risks involving the inaccuracies of AI psychotherapy models, it is unclear where the consequences should be felt in the case of an accident. Questions of liability extend across the organizations that deploy these systems, creating ambiguity in accountability. Inaccurate or misleading outputs may delay appropriate treatment or reinforce harmful beliefs. Data privacy concerns are heightened by the sensitive nature of mental health information and the potential for misuse or breaches. There is also a risk of overreliance on automated systems, which may reduce critical human oversight in complex clinical situations. Existing regulatory frameworks have not fully adapted to address these challenges, leaving gaps in standards and enforcement. Transparency around model limitations and decision-making processes remains inconsistent. These factors collectively underscore the need for clearer guidelines, stronger safeguards, and defined accountability structures. It would be beneficial to explore or create policies such as liability frameworks, clinical oversight requirements, privacy regulations, and crisis escalation protocols to protect individuals who utilize these tools and ensure steps are being taken to oversee data and privacy concerns. Much farther into the future, assuming AI and psychotherapy become increasingly amalgamated, it is paramount to assess the long-term physical, mental, and psychological strain that is associated with the overuse of AI. (Rahsepar Meadi, M. et al.), a scoping review, warns against overuse as it can alter human emotion and belief system, and can lead to an unhealthy detachment from oneself. Additionally, to minimise the risk of inaccuracy in emergency situations, models should be trained to flag self-harming or suicidal thoughts from the user and delegate those situations to trained professionals. These considerations suggest that the integration of AI into psychotherapy must be accompanied by clear usage boundaries and ongoing monitoring of user

outcomes. Without such safeguards, the scalability of these tools could amplify harm rather than mitigate it, particularly among vulnerable populations.

CONCLUSION

AI presents a solution to special cases of mental health issues by making it a more accessible and less stigmatized version of current mental health services. As artificial intelligence continues to reshape psychiatric care, its design, implementation, and evaluation must be firmly grounded in ethical standards that uphold human dignity, promote well-being, and ensure fairness. Their effectiveness ultimately depends on careful oversight, thorough validation, and interdisciplinary cooperation. Establishing systems that emphasize equity, safety, and clinical reliability is essential to ensuring that technological advancements translate into meaningful improvements in care. Only through this approach can AI be responsibly leveraged to enhance mental health outcomes on a global scale.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Dr. Dan-Dan Li, for her endless support and guidance through this journey. I greatly value all the feedback I received from her for this paper, and every other piece of advice she gave along the way.

REFERENCES

Amabili, G. et al. “Managing Cognitive Decline Through a Social Robot-Based Intervention: Protocol for the engAGE Proof of Concept and Randomized Controlled Trial.” *JMIR Research Protocols*, vol. 14, 2025, pp. e67601, doi: 10.2196/67601.

Andersson, G. “The latest developments with internet-based psychological treatments for depression.” *Expert Review of Neurotherapeutics*, vol. 24, no. 2, 2024, pp. 171-176, doi: 10.1080/14737175.2024.2309237.

Boit, S., and R. Patil. “A Prompt Engineering Framework for Large Language Model-Based Mental Health Chatbots: Conceptual Framework.” *JMIR Mental Health*, vol. 12, 2025, pp. e75078, doi: 10.2196/75078.

Bucher, A. et al. “"It's Not Only Attention We Need": Systematic Review of Large Language Models in Mental Health Care.” *JMIR Mental Health*, vol. 12, 2025, pp. e78410, doi: 10.2196/78410.

Cabrera Lozoya, D. et al. “Leveraging Large Language Models for Simulated Psychotherapy Client Interactions: Development and Usability Study of Client101.” *JMIR Medical Education*, vol. 11, 2025, pp. e68056, doi: 10.2196/68056.

Campellone, T. R. et al. "Safety and User Experience of a Generative Artificial Intelligence Digital Mental Health Intervention: Exploratory Randomized Controlled Trial." *Journal of Medical Internet Research*, vol. 27, 2025, pp. e67365, doi: 10.2196/67365.

Cheng, S. W. et al. "The now and future of ChatGPT and GPT in psychiatry." *Psychiatry and Clinical Neurosciences*, vol. 77, no. 11, 2023, pp. 592-596, doi: 10.1111/pcn.13588.

Chiu, Y. H. et al. "Using cognitive behavioral therapy-based chatbots to alleviate symptoms of insomnia, depression, and anxiety: A randomized controlled trial." *Health Informatics Journal*, vol. 31, no. 4, 2025, pp. 14604582251396428, doi: 10.1177/14604582251396428.

Dallison, S. et al. "Using Generative AI to Co-Design Digital Mental Health Interventions With Adolescents in Rural South Africa: Qualitative Thematic Analysis of Participatory Workshops." *Journal of Medical Internet Research*, vol. 27, 2025, pp. e73535, doi: 10.2196/73535.

de Graaff, A. M. et al. "Evaluation of a Guided Chatbot Intervention for Young People in Jordan: Feasibility Randomized Controlled Trial." *JMIR Mental Health*, vol. 12, 2025, pp. e63515, doi: 10.2196/63515.

Feng, Y. et al. "Effectiveness of AI-Driven Conversational Agents in Improving Mental Health Among Young People: Systematic Review and Meta-Analysis." *Journal of Medical Internet Research*, vol. 27, 2025, pp. e69639, doi: 10.2196/69639.

Fietta, V. et al. "Chatbot-Based Version of a World Health Organization-Validated Intervention for Stress Management in Patients With Breast Cancer (Self-Help Plus): Protocol for a Pilot Feasibility Study." *JMIR Research Protocols*, vol. 14, 2025, pp. e65837, doi: 10.2196/65837.

Fietta, V. et al. "World Health Organization Evidence-Based Self-Help Plus Intervention for Stress Management via Chatbot: Protocol for Adaptation to a Tech-Enabled Model." *JMIR Research Protocols*, vol. 14, 2025, pp. e69644, doi: 10.2196/69644.

Gkintoni, E. et al. "Next-Generation Cognitive-Behavioral Therapy for Depression: Integrating Digital Tools, Teletherapy, and Personalization for Enhanced Mental Health Outcomes." *Medicina*, vol. 61, no. 3, 2025, pp. 431, doi: 10.3390/medicina61030431.

Guglielmucci, F., and D. Di Basilio. "Predicting Engagement With Conversational Agents in Mental Health Therapy by Examining the Role of Epistemic Trust, Personality, and Fear of Intimacy: Cross-Sectional Web-Based Survey Study." *JMIR Human Factors*, vol. 12, 2025, pp. e70698, doi: 10.2196/70698.

Guhan, P. et al. "Developing a Machine Learning-Based Automated Patient Engagement Estimator for Telehealth: Algorithm Development and Validation Study." *JMIR Formative Research*, vol. 9, 2025, pp. e46390, doi: 10.2196/46390.

Haber, Y. et al. "The Artificial Third: A Broad View of the Effects of Introducing Generative Artificial Intelligence on Psychotherapy." *JMIR Mental Health*, vol. 11, 2024, pp. e54781, doi: 10.2196/54781.

Habicht, J. et al. "Generative AI-Enabled Therapy Support Tool for Improved Clinical Outcomes and Patient Engagement in Group Therapy: Real-World Observational Study." *Journal of Medical Internet Research*, vol. 27, 2025, pp. e60435, doi: 10.2196/60435.

Haupt, M. R. et al. "Evaluating the Influence of Role-Playing Prompts on ChatGPT's Misinformation Detection Accuracy: Quantitative Study." *JMIR Infodemiology*, vol. 4, 2024, pp. e60678, doi: 10.2196/60678.

Held, P. et al. "AI-Facilitated Cognitive Reappraisal via Socrates 2.0: Mixed Methods Feasibility Study." *JMIR Mental Health*, vol. 12, 2025, pp. e80461, doi: 10.2196/80461.

Hong, J. et al. "Randomised controlled trial of VR-based observation meditation with AI coaching ('Otti') for stress reduction in university students in the United States: study protocol." *BMJ Open*, vol. 15, no. 10, 2025, pp. e097236, doi: 10.1136/bmjopen-2024-097236.

Hudon, A. et al. "Navigating the Maze of Social Media Disinformation on Psychiatric Illness and Charting Paths to Reliable Information for Mental Health Professionals: Observational Study of TikTok Videos." *Journal of Medical Internet Research*, vol. 27, 2025, pp. e64225, doi: 10.2196/64225.

Immel, D. et al. "Patients' and Health Care Professionals' Expectations of Virtual Therapeutic Agents in Outpatient Aftercare: Qualitative Survey Study." *JMIR Formative Research*, vol. 9, 2025, pp. e59527, doi: 10.2196/59527.

Jabir, A. I. et al. "Attrition in Conversational Agent-Delivered Mental Health Interventions: Systematic Review and Meta-Analysis." *Journal of Medical Internet Research*, vol. 26, 2024, pp. e48168, doi: 10.2196/48168.

Jabir, A. I. et al. "Evaluating Conversational Agents for Mental Health: Scoping Review of Outcomes and Outcome Measurement Instruments." Journal of Medical Internet Research, vol. 25, 2023, pp. e44548, doi: 10.2196/44548.

Jesudason, D. et al. "Artificial intelligence (AI) in psychotherapy: A challenging frontier." *Australasian Psychiatry*, vol. 33, no. 4, 2025, pp. 629-632, doi: 10.1177/10398562251346075.

Jurblum, M., and R. Selzer. "Potential promises and perils of artificial intelligence in psychotherapy -The AI Psychotherapist (APT)." *Australasian Psychiatry*, vol. 33, no. 1, 2025, pp. 103-105, doi: 10.1177/10398562241286312.

Kuhlmeier, F. O. et al. "Designing Chatbots to Treat Depression in Youth: Qualitative Study." *JMIR Human Factors*, vol. 12, 2025, pp. e66632, doi: 10.2196/66632.

Li, S. et al. "Determinants of rural middle school students' adoption of AI chatbots for mental health." *Frontiers in Public Health*, vol. 13, 2025, pp. 1619535, doi: 10.3389/fpubh.2025.1619535.

Malouin-Lachance, A. et al. "Does the Digital Therapeutic Alliance Exist? Integrative Review." *JMIR Mental Health*, vol. 12, 2025, pp. e69294, doi: 10.2196/69294.

Napiwotzki, I. et al. "Comparing Human and AI Therapists in Behavioral Activation for Depression: Cross-Sectional Questionnaire Study." *JMIR Formative Research*, vol. 9, 2025, pp. e78138, doi: 10.2196/78138.

Palmer, C. E. "Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care." *Journal of Medical Internet Research*, vol. 27, 2025, pp. e69351, doi: 10.2196/69351.

Poudel, U. et al. "AI in Mental Health: A Review of Technological Advancements and Ethical Issues in Psychiatry." *Issues in Mental Health Nursing*, vol. 46, no. 7, 2025, pp. 693-701, doi: 10.1080/01612840.2025.2502943.

Prasad, N. et al. "Deep learning for the prediction of clinical outcomes in internet-delivered CBT for depression and anxiety." *PLOS One*, vol. 18, no. 11, 2023, pp. e0272685, doi: 10.1371/journal.pone.0272685.

Rahsepar Meadi, M. et al. "Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review." *JMIR Mental Health*, vol. 12, 2025, pp. e60432, doi: 10.2196/60432.

Reyes-Portillo, J. A. et al. "Generative AI-Powered Mental Wellness Chatbot for College Student Mental Wellness: Open Trial." *JMIR Formative Research*, vol. 9, 2025, pp. e71923, doi: 10.2196/71923.

Roessner, V. et al. "Taming the chaos ?! - part 2: limitations of using eXplainable Artificial Intelligence (XAI) to tackle the complexity in psychotherapy of children and adolescents." *European Child & Adolescent Psychiatry*, vol. 33, no. 6, 2024, pp. 1617-1619, doi: 10.1007/s00787-024-02487-7.

Rubin, M. et al. "Considering the Role of Human Empathy in AI-Driven Therapy." *JMIR Mental Health*, vol. 11, 2024, pp. e56529, doi: 10.2196/56529.

Sadeh-Sharvit, S. et al. "Effects of an Artificial Intelligence Platform for Behavioral Interventions on Depression and Anxiety Symptoms: Randomized Clinical Trial." *Journal of Medical Internet Research*, vol. 25, 2023, pp. e46781, doi: 10.2196/46781.

Schillings, C. et al. "Effects of a Chatbot-Based Intervention on Stress and Health-Related Parameters in a Stressed Sample: Randomized Controlled Trial." *JMIR Mental Health*, vol. 11, 2024, pp. e50454, doi: 10.2196/50454.

Scholich, T. et al. "A Comparison of Responses from Human Therapists and Large Language Model-Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study." *JMIR Mental Health*, vol. 12, 2025, pp. e69709, doi: 10.2196/69709.

Six, S. et al. "Impact of Conversational and Animation Features of a Mental Health App Virtual Agent on Depressive Symptoms and User Experience Among College Students: Randomized Controlled Trial." *JMIR Mental Health*, vol. 12, 2025, pp. e67381, doi: 10.2196/67381.

Sobowale, K. et al. "Evaluating Generative AI Psychotherapy Chatbots Used by Youth: Cross-Sectional Study." *JMIR Mental Health*, vol. 12, 2025, pp. e79838, doi: 10.2196/79838.

Sobowale, K., and D. K. Humphrey. "Evaluating the Quality of Psychotherapy Conversational Agents: Framework Development and Cross-Sectional Study." *JMIR Formative Research*, vol. 9, 2025, pp. e65605, doi: 10.2196/65605.

Spytska, L. "The use of artificial intelligence in psychotherapy: development of intelligent therapeutic systems." *BMC Psychology*, vol. 13, no. 1, 2025, pp. 175, doi: 10.1186/s40359-025-02491-9.

Stephenson, C. et al. "Evaluation of an Artificial Intelligence and Online Psychotherapy Initiative to Improve Access and Efficiency in an Ambulatory Psychiatric Setting: Évaluation d'une initiative de psychothérapie en ligne basée sur l'intelligence artificielle visant à améliorer l'accès et l'efficacité en milieu psychiatrique ambulatoire." *Canadian Journal of Psychiatry*, vol. 71, no. 1, 2026, pp. 30-40, doi: 10.1177/07067437251355641.

Sun, X. et al. "Use of AI-Based Mental Health Tools and Psychological Well-Being Among Chinese University Students: A Parallel Mediation Model of Emotional Self-Efficacy and Perceived Autonomy." *Scientific Reports*, vol. 15, no. 1, 2025, pp. 40305, doi: 10.1038/s41598-025-24013-8.

Szoke, D. et al. "Artificial Intelligence in Mental Health Services Under Illinois Public Act 104-0054: Legal Boundaries and a Framework for Establishing Safe, Effective AI Tools." *JMIR Mental Health*, vol. 12, 2025, pp. e84854, doi: 10.2196/84854.

Tabassum, A. et al. "Exploring the Application of AI and Extended Reality Technologies in Metaverse-Driven Mental Health Solutions: Scoping Review." *Journal of Medical Internet Research*, vol. 27, 2025, pp. e72400, doi: 10.2196/72400.

Thunström, A. O. et al. "Behavior Emotion Therapy System and You: Co-Design and Evaluation of a Mental Health Chatbot and Digital Human for Mild to Moderate Anxiety in Healthy Participants." *JMIR Formative Research*, vol. 9, 2025, pp. e66163, doi: 10.2196/66163.

Tong, A. C. Y. et al. "Effectiveness of Topic-Based Chatbots on Mental Health Self-Care and Mental Well-Being: Randomized Controlled Trial." *Journal of Medical Internet Research*, vol. 27, 2025, pp. e70436, doi: 10.2196/70436.

van der Schyff, E. L. et al. "Providing Self-Led Mental Health Support Through an Artificial Intelligence-Powered Chat Bot (Leora) to Meet the Demand of Mental Health Care." *Journal of Medical Internet Research*, vol. 25, 2023, pp. e46448, doi: 10.2196/46448.

Wang, L. et al. "Evaluating Generative AI in Mental Health: Systematic Review of Capabilities and Limitations." *JMIR Mental Health*, vol. 12, 2025, pp. e70014, doi: 10.2196/70014.

Xu, Z. et al. "The Digital Therapeutic Alliance With Mental Health Chatbots: Diary Study and Thematic Analysis." *JMIR Mental Health*, vol. 12, 2025, pp. e76642, doi: 10.2196/76642.

Yang, Y. et al. "A New Research Model for Artificial Intelligence-Based Well-Being Chatbot Engagement: Survey Study." *JMIR Human Factors*, vol. 11, 2024, pp. e59908, doi: 10.2196/59908.

Zhang, X. et al. "Exploring Body Image Awareness With a Large Language Model-Based Conversational Agent: Qualitative Study With Young Adults." *Journal of Medical Internet Research*, vol. 27, 2025, pp. e78829, doi: 10.2196/78829.

ABOUT THE AUTHOR

Shruti Shyam is a junior at Northgate High School in Walnut Creek, California. She hopes to pursue a career in healthcare research, specifically pharmacology and immunology. She is also interested in artificial intelligence's implications on modern society, especially in how it will develop modern healthcare and drive scientific discovery.