RESEARCH REPORT

Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review

Yining Hua[1,2], Steve Siddals[2], Zilin Ma[3], Isaac Galatzer-Levy[4,5], Winna Xia[2], Christine Hau[2], Hongbin Na[6], Matthew Flathers[2], Jake Linardon[7], Cyrus Ayubcha[7], John Torous[2]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA; [2]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA; [3]Intelligent Interactive Systems Group, Harvard School of Engineering and Applied Sciences, Allston, MA, USA; [4]Department of Psychiatry, New York University Grossman School of Medicine, New York, NY, USA; [5]Google Research, New York, NY, USA; [6]Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, NSW, Australia; [7]SEED Lifespan Strategic Research Centre, School of Psychology, Faculty of Health, Deakin University, Geelong, VIC, Australia

The rapid evolution of artificial intelligence (AI) chatbots in mental health care presents a fragmented landscape with variable clinical evidence and evaluation rigor. This systematic review of 160 studies (2020-2024) classifies chatbot architectures – rule-based, machine learning-based, and large language model (LLM)-based – and proposes a three-tier evaluation framework: foundational bench testing (technical validation), pilot feasibility testing (user engagement), and clinical efficacy testing (symptom reduction). While rule-based systems dominated until 2023, LLM-based chatbots surged to 45% of new studies in 2024. However, only 16% of LLM studies underwent clinical efficacy testing, with most (77%) still in early validation. Overall, only 47% of studies focused on clinical efficacy testing, exposing a critical gap in robust validation of therapeutic benefit. Discrepancies emerged between marketed claims ("AI-powered") and actual AI architectures, with many interventions relying on simple rule-based scripts. LLM-based chatbots are increasingly studied for emotional support and psychoeducation, yet they pose unique ethical concerns, including incorrect responses, privacy risks, and unverified therapeutic effects. Despite their generative capabilities, LLMs remain largely untested in high-stakes mental health contexts. This paper emphasizes the need for standardized evaluation and benchmarking aligned with medical AI certification to ensure safe, transparent and ethical deployment. The proposed framework enables clearer distinctions between technical novelty and clinical efficacy, offering clinicians, researchers and regulators ordered steps to guide future standards and benchmarks. To ensure that AI chatbots enhance mental health care, future research must prioritize rigorous clinical efficacy trials, transparent architecture reporting, and evaluations that reflect real-world impact rather than the well-known potential.

Key words: Artificial intelligence, chatbots, rule-based systems, machine learning, large language models, foundational bench testing, pilot feasibility testing, clinical efficacy testing, mental health care

(World Psychiatry 2025;24:383–394)

Mental disorders remain a major contributor to the global burden of disease. An estimated 970 million individuals live with mental health or substance use disorders worldwide, with depression and anxiety among the leading causes of disability-adjusted life years lost[1,2].

Despite the increasing recognition of mental health as a critical component of public health, care systems remain underfunded and overwhelmed[3], with fewer than five mental health professionals available per 100,000 people globally[4]. This unmet need is more pronounced in low- and middle-income countries, where more than 75% of individuals with mental health conditions fail to receive treatment[5]. The need for scalable, accessible solutions has driven interest in digital interventions, particularly conversational agents or chatbots, as potential tools to support mental health care by way of screening, psychoeducation, and therapy augmentation[6].

While the rapid growth of artificial intelligence (AI)-driven chatbots offers new possibilities for mental health applications, their potential remains unclear. Many clinicians, policy makers and researchers are uncertain about whether these tools are appropriate for clinical deployment. Much of this uncertainty arises from the heterogeneous nature of chatbot technologies, which range from basic rule-based systems to advanced large language models (LLMs). While rule-based chatbots rely on pre-programmed scripts or decision trees, LLMs leverage deep neural networks trained on vast datasets to produce more versatile, human-like conversational capabilities.

The conflation of these technologies has led to several challenges. While most health care chatbots remain rule-based, companies often continue to market them as "AI" or even "LLM-driven". This creates misconceptions about their sophistication and reliability, as these systems differ substantially in their capabilities and limitations. Additionally, exaggerated claims about LLMs, such as their ability to pass professional exams or display empathy surpassing that of human clinicians, have blurred the distinction between experimental results and real-world applicability[7,8]. For instance, passing a written test or simulating empathetic dialogue in controlled conditions does not necessarily equate to making accurate clinical diagnoses or supporting patients in complex and dynamic clinical scenarios.

The current state of the AI chatbot literature reflects these challenges. Meta-analyses and reviews often conflate simple feasibility testing results with more complex clinically focused research. In this paper, we applied a staged approach to translational research and clinical development that mirrors the progression from basic science/preclinical research to early phase human testing (phase 1 or 2 clinical trials) to clinical efficacy testing (phase 3 clinical trials). This scheme organizes chatbot studies into three tiers: foundational bench testing for technical feasibility, pilot feasibility testing for assessment of usability and acceptability in humans, and clinical efficacy testing. These categories reflect increasing levels of clinical applicability, providing a structured approach to understanding the field's progress. This approach contextualizes the

current discussion around AI certification with a staged structure[9], and provides a useful scaffold for understanding prior studies and developing future work.

The evolution of AI chatbots reflects decades of advancements in computational approaches to human language, which we categorize into three paradigms: rule-based systems, non-LLM machine learning models, and LLM-based systems. This tripartite framework is designed to clarify distinctions between these paradigms, particularly as the term "AI" has historically encompassed a wide range of technologies, from early deterministic systems to modern probabilistic models.

Rule-based systems were the earliest form of conversational AI. ELIZA, developed in the 1960s, exemplified these systems by simulating a Rogerian psychotherapist through simple pattern-matching and substitution rules[10]. Currently popular mental health chatbots, such as *Woebot*, have been primarily rule-based systems, underscoring the enduring effectiveness of this approach[11]. Most chatbots continue to be employed in narrowly defined tasks, such as structured screening tools or symptom checkers[12], where deterministic outputs remain sufficient. However, their inability to adapt to novel inputs or provide personalized responses has constrained their utility where dynamic and context-sensitive interactions are critical[13]. Despite limitations, rule-based systems laid the groundwork for subsequent innovations by demonstrating the feasibility of automated dialogue.

The transition from rule-based systems to machine learning introduced greater adaptability and probabilistic reasoning into chatbot designs. Machine learning refers to algorithms that identify patterns in data, enabling models to generalize beyond explicit rules and make predictions based on statistical likelihoods[14,15]. Non-LLM machine learning-based chatbots marked a pivotal shift by moving beyond scripted interactions. These systems incorporated natural language processing techniques, such as sentiment analysis and intent recognition, to infer user emotions and tailor responses to context[16]. For example, conversational agents such as *Wysa* employ a combination of machine learning algorithms and rule-based scripts to deliver cognitive-behavioral therapy (CBT) interventions, demonstrating efficacy in reducing symptoms of depression and anxiety in pilot studies[17,18].

While non-LLM machine learning models expanded the scope of chatbot applications, they faced inherent challenges. Their performance often depends on domain-specific training data, which limits generalizability across diverse conversational scenarios. Additionally, these systems struggle with generating coherent and human-like language, as they were typically designed to classify or process inputs rather than produce contextually appropriate outputs.

LLMs represent a paradigm shift in AI chatbots, driven by their ability to generate human-like language with fluency and contextual awareness. Built on Transformer architectures, LLMs such as OpenAI's *GPT* series and Meta's *Llama* models leverage self-attention mechanisms to understand relationships between words and concepts across extensive text passages[19,20]. This architecture enables LLMs to maintain coherence within complex, multi-turn dialogues in ways that previous approaches could not achieve[21]. Unlike earlier machine learning-based systems that primarily focused on classification or limited response selection, LLMs are fundamentally generative in nature – they create novel text rather than selecting from predefined responses.

LLMs' advanced linguistic capabilities have sparked explosive interest in their mental health applications[22], offering the potential to enhance psychoeducation, triage, and supportive interactions. However, the transformative potential of LLMs is accompanied by significant challenges. Their reliance on vast, uncurated datasets introduces risks such as bias, misinformation, and the generation of fabricated or harmful content[23]. In psychiatry, where the consequences of errors can be severe, these limitations raise ethical concerns about reliability and safety.

METHODS

We classified chatbots into three systems:

- *Rule-based systems*. These rely on deterministic scripts (e.g., rule-based conversation systems, simple decision trees), with no data-driven learning. They are ideal for structured, low-risk tasks (e.g., symptom checklists) where predictability ensures safety. However, their rigidity limits their utility in dynamic therapeutic contexts.
- *Machine learning-based systems*. These include traditional machine learning (e.g., support vector machine, SVM) and non-generative deep learning (e.g., recurrent neural networks, RNN; long short-term memory, LSTM; and bidirectional encoder representations from transformers, BERT). While RNN/LSTM and traditional machine learning differ technically (e.g., sequential vs. static data processing), both lack natural language fluency. Grouping these under "machine learning-based systems" reflects their shared limitation in mental health: adaptability without generative capacity.
- *LLM-based systems*. These leverage generative models trained on vast text corpora to produce human-like dialogue. This category includes multimodal models that can process images, audio or other modalities in addition to text, as long as they maintain the core LLM architecture for language generation.

We used a tiered framework that categorizes studies by their evaluation rigor, akin to the translational pipeline from technical validation to real-world clinical impact:

- *T1. Foundational bench testing*. This focuses on technical validation in controlled settings (e.g., scripted scenarios, expert assessments) to ensure that chatbots meet baseline functional and safety standards. For mental health, this stage is critical to verify adherence to clinical guidelines (e.g., suicide risk protocols) before human interaction.
- *T2. Pilot feasibility testing*. This assesses usability and acceptability with human participants (e.g., patients, clinicians) over short-term interactions. While it provides insights into engagement, it often overlooks sustained therapeutic outcomes – a

gap particularly problematic in mental health, where longitudinal efficacy is paramount.

- *T3. Clinical efficacy testing*. This measures clinically meaningful outcomes (e.g., symptom reduction via validated rating scales) over extended periods. It is essential for mental health chatbots, as transient usability gains (T2) do not necessarily equate to therapeutic benefit.

This framework ensures that mental health interventions undergo rigorous validation before clinical deployment. A chatbot that performs well in scripted tests (T1) may still fail in real-world empathy or crisis management, while short-term usability (T2) does not guarantee long-term adherence or relapse prevention. By stratifying evidence into three tiers, the classification enables clinicians and regulators to distinguish technically functional tools from those with proven clinical impact[23].

We systematically reviewed mental health chatbot studies published from January 1, 2020 to January 1, 2025 across PubMed, APA PsycNet, Scopus and Web of Science, following PRISMA 2020 guidelines[24]. Search strings were adapted from prior scoping reviews and optimized for lexical coverage (see supplementary information). Additional records were identified through manual searches of Google Scholar and major AI conference proceedings.

To ensure relevance, only studies evaluating chatbots within a mental health care context were included. Papers focused on psycholinguistics, psychosocial demographics, or predictive models without conversational interfaces were excluded. Only full, peer-reviewed papers written in English were considered. Reviews, meta-analyses and retracted papers were excluded. Eligible studies were required to include an actual evaluation of chatbot performance, excluding protocol descriptions.

The screening and annotation process was a collaborative effort involving the entire research team. Each study was randomly assigned to at least two reviewers, who independently evaluated its eligibility based on predefined inclusion criteria. Any disagreements between reviewers were resolved through group discussions to ensure consistency.

Figure 1 presents the PRISMA flow diagram illustrating the study selection process. A total of 1,727 records were identified through database and manual searches, including 620 from PubMed, 18 from APA PsycNet, 419 from Scopus, 480 from Web of Science, 131 from Google Scholar, and 59 from major AI conferences. After removing 790 duplicates, 937 unique records were screened. Of these, 734 studies were excluded based on title and abstract screening, due to irrelevance or failure to meet the inclusion criteria. Following this, 203 reports were sought for full-text retrieval, but four could not be retrieved. The remaining 199 full-text reports were assessed for eligibility, leading to the exclusion of eight studies that lacked chatbot evaluation, 21 that were duplicates published under different titles, and ten whose models did not meet our definition of LLMs. Ultimately, 160 studies met the inclusion criteria and were included in the systematic review[25-184].

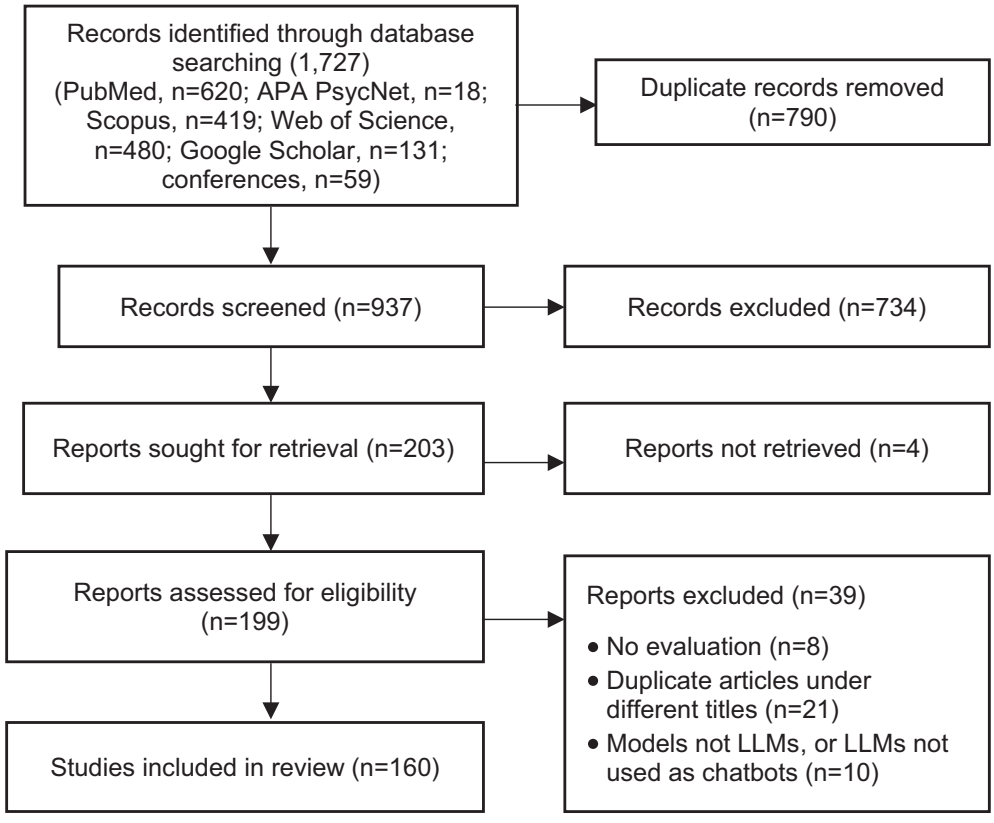

Figure 1 PRISMA flow diagram of study selection process. LLMs – large language models

To ensure consistency and depth in data extraction, senior team members (YH and JT) implemented a structured annotation protocol (see supplementary information), formalized through team training. Key elements included chatbot architecture, evaluation methodology, target conditions, functional purpose, and outcome measures. Each study was also annotated with type-specific information based on its evaluation tier (T1, T2 or T3), capturing evaluator characteristics, usage duration, and relevant clinical instruments. Missing or non-applicable data were systematically flagged to support transparent synthesis.

To analyze the 160 annotated studies, we implemented a structured multi-step methodology to transform raw annotations into standardized themes. In cases where studies were classified into multiple categories, each classification was weighted proportionally (e.g., a study targeting both depression and anxiety would be counted twice, each with weighting 0.5). Reported subtotals and percentages were rounded to the nearest whole number, which may result in apparent summation discrepancies. Computational tools supported initial categorization, with all results refined and validated by domain experts to ensure fidelity and interpretability (see also supplementary information).

RESULTS

Evolution of chatbot architectures

Research interest in mental health chatbots increased substantially over the review period, with the annual number of studies quadrupling from 14 in 2020 to 56 in 2024. Coinciding with this growth, the underlying chatbot architectures underwent a significant transformation (see Figure 2).

Initial research in 2020 (n=14)[25-38] focused exclusively on rule-based systems (100%). The landscape diversified from 2021 (n=28)[39-66], with the emergence of machine learning-based (21%) and the first LLM-based studies (11%), although rule-based systems remained dominant (68%). Machine learning-based approaches peaked in 2022 (40% of 25 studies[67-91]), before stabilizing as a smaller component (14% in 2024).

LLM-based architectures, after comprising 16% of studies in 2022 and 19% in 2023 (n=37)[92-128], surged to represent 45% of studies in 2024 (n=56)[129-184]. This rapid rise made LLMs the most frequently studied architecture in 2024, surpassing rule-based systems (41%), despite the absolute number of rule-based studies remaining relatively consistent in 2023-2024. This trend indicates a decisive shift towards investigating advanced generative models within the field.

Distribution by evaluation methodology and architecture

Research effort across the evaluation stages was primarily focused on clinical efficacy testing (n=75), followed by pilot feasibility testing (n=72), and foundational bench testing (n=13) (see Figure 3).

Chatbot architecture distribution varied markedly across these evaluation stages. Foundational bench testing was dominated by LLM-based systems, which accounted for over two-thirds (77%) of studies at this stage, with smaller contributions from machine learning-based (15%) and rule-based (8%) systems. Rule-based architectures predominated in later stages, accounting for over half of both pilot feasibility studies (58%) and clinical efficacy trials (65%). The proportion of LLM-based studies decreased substantially in these stages, accounting for only 24% of pilot feasibility studies and 16% of clinical efficacy studies. Machine learning-based systems remained a minority across all stages, ranging from 15% in foundational bench testing to 19% in clinical efficacy testing.

This stark contrast between stages indicates that, while LLMs are the primary focus of early technical validation, rule-based systems remain the principal architecture undergoing human testing and clinical evaluation.

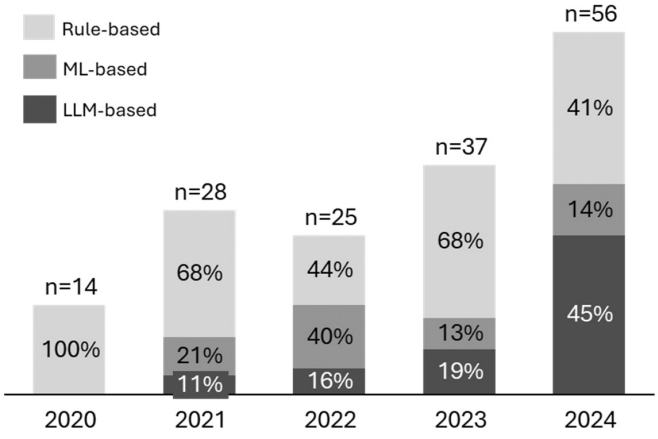

Figure 2 Evolution of chatbot architectures studied from 2020 to 2024. Percentages indicate the proportion of studies utilizing each architecture type within a given year. The number above each bar indicates the total number of studies for that year. ML – machine learning, LLM – large language model.

Target conditions, functional purpose, and outcome measures of chatbot studies

Analysis of the studies' target conditions, functional purpose, and outcome measures revealed distinct patterns in the application and evaluation of different chatbot architectures (see Figure 4).

Examining target conditions, research most frequently addressed general mental well-being (n=51), depression (n=50), and anxiety (n=41). Rule-based systems were the predominant architecture for studies targeting depression (58%) and anxiety (62%), compared to LLM (20-25%) and machine learning-based (13-23%) systems. LLM-based systems showed higher relative representation in studies targeting general mental well-being (28%), compared with rule-based (49%) and machine learning-based (22%) approaches.

Methodologically, studies of general mental well-being were largely in pilot feasibility testing (66%), with fewer in clinical efficacy testing (27%) or foundational bench testing (7%). By contrast, both depression and anxiety interventions had mostly advanced

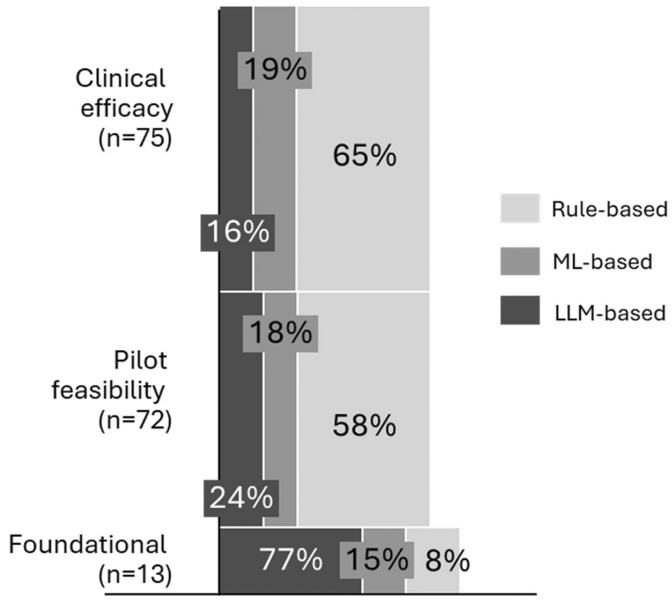

Figure 3 Distribution of studies by evaluation methodology and chatbot architecture. Percentages indicate the proportion of chatbot architectures within each evaluation methodology category. Subtotals and percentages are rounded to the nearest whole number, which may result in apparent summation discrepancies. ML – machine learning, LLM – large language model.

to clinical efficacy trials (57% and 58%, respectively), with smaller proportions in pilot feasibility (38% for depression; 33% for anxiety) and foundational testing (6-8%).

Studies were grouped into five functional purposes, including emotional support (n=46), therapeutic interventions (n=42), education and skills training (n=27), assessment and monitoring (n=21), and general and other functions (n=24). In every category, rule-based systems dominated, ranging from 48% in emotional support to 83% in general and other functions. LLM-based approaches were most represented in emotional support (30%) and assessment and monitoring (29%), while machine learning-based systems peaked at 27% for therapeutic interventions and dipped as low as 4% for general and other functions.

Evaluation methodology varied notably by functional purpose. Pilot feasibility testing was the most common method for assessment and monitoring (78%), emotional support (53%), and education and skills training (49%). By contrast, clinical efficacy testing led in therapeutic interventions (65%), and general and other functions (71%), with education and skills (43%) and emotional support (34%) also seeing substantial clinical work. Foundational bench testing remained minimal, accounting for 7-13% of studies in four categories and 0% for general and other studies.

The choice of chatbot architecture and evaluation stage was associated with the outcome measures prioritized. Studies measuring clinical outcomes (n=99) predominantly employed rule-based (59%) or machine learning-based (18%) systems, with fewer using LLM-based approaches (23%). These clinical outcome studies were most often evaluated via clinical efficacy testing (65%), followed by pilot feasibility (31%) and foundational bench testing (4%).

User experience evaluations (n=46) similarly favored rule-based (64%) and machine learning-based (19%) architectures over LLM-based approaches (17%), and were overwhelmingly conducted as pilot feasibility studies (78%), with smaller proportions in clinical efficacy (17%) and foundational testing (5%). In contrast, technical performance studies (n=15) were dominated by LLM-based systems (54%), with rule-based and machine learning-based approaches at 31% and 15% respectively, and were primarily assessed at the foundational bench (45%) and pilot feasibility (37%) stages, with only 18% reaching clinical efficacy testing.

Characteristics of evaluation stages

The nature of evaluations conducted on mental health chatbots evolved significantly across the research pipeline, particularly concerning the types of participants involved (see Figure 5).

Foundational bench testing (n=13) primarily involved evaluations conducted with clinicians (62%) or represented technical assessments where participant type was "not applicable" (38%). Transitioning to the pilot feasibility testing (n=72), evaluation efforts focused predominantly on general users (78%) to assess usability and acceptability, alongside a substantial inclusion of patients (19%) for initial target population testing. The clinical efficacy testing (n=75) presented a more varied participant profile, characterized by the engagement of clinicians as evaluators (19%) and continued recruitment of general users (25%). Notably, explicitly identified patient participants were less frequently involved (5%) in this final stage compared to pilot studies, while a large proportion of these efficacy trials reported participant type as "not applicable" (37%) or "not specified" (13%), potentially reflecting diverse study designs or reporting practices.

During the pilot feasibility phase (n=72), study durations ranged from under one hour to two years (see Figure 6). Under one hour applied to eight studies, between one hour and one day to eighteen studies, and between one day and one week to sixteen studies. Ten studies did not report a duration. Rule-based architectures appeared in half to two-thirds of studies across every duration band. LLM-based systems featured in 20% to 33% of studies, with a 31% share in the one-day to one-week group. Machine learning-based approaches ranged from 0% in the not-applicable category up to 25% in the one-to-four-week interval. These results show that, although rule-based chatbots dominate pilot feasibility testing, LLM and machine learning systems have also been trialed across the full spectrum of study lengths.

Terminology discrepancies: the meaning of "AI"

Beyond the core findings related to chatbot architectures and evaluation stages, the terminology used to describe these technologies in study titles also warranted examination (see Figure 7).

Analysis revealed ambiguity in the use of the term "AI". Of the 160 included studies, a small proportion (n=21, 13%) explicitly

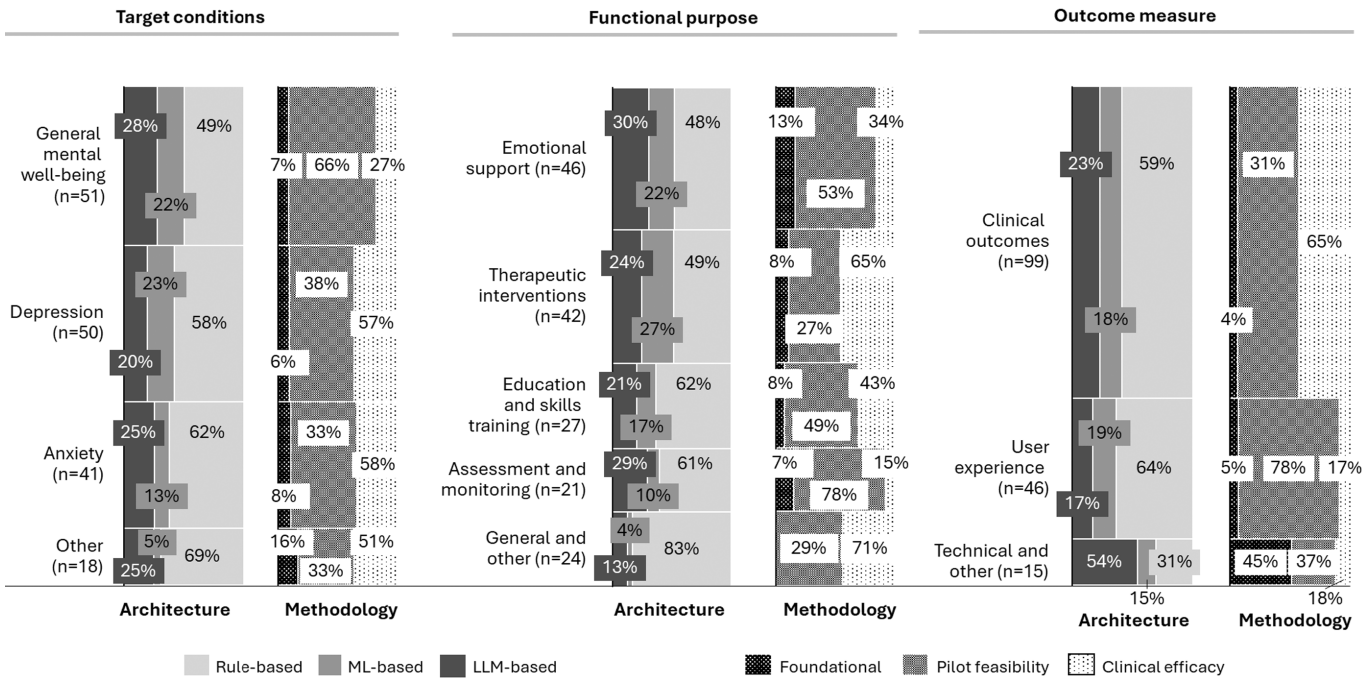

Figure 4 Distribution of chatbot studies by target condition, functional purpose, and outcome measure. Within each subcategory, the left bar indicates the percentage distribution of chatbot architectures used, and the right bar shows the percentage distribution of evaluation methodologies employed. Subtotals and percentages are rounded to the nearest whole number, which may result in apparent summation discrepancies. ML – machine learning, LLM – large language model.

used the term "AI" in their titles; the vast majority (n=139, 87%) did not. Among the 21 studies that did use the "AI" label, the majority (57%) employed advanced LLM-based systems. A further 19% utilized machine learning-based approaches. Notably, however, the "AI" label was also applied in nearly one-quarter (24%) of these cases to studies utilizing rule-based architectures. As rule-based systems operate on predefined scripts without the adaptive learning capabilities characteristic of contemporary machine learning and LLM systems, their inclusion under the general "AI" descriptor contributes to terminological ambiguity and potential misrepresentation of chatbot sophistication within the field.

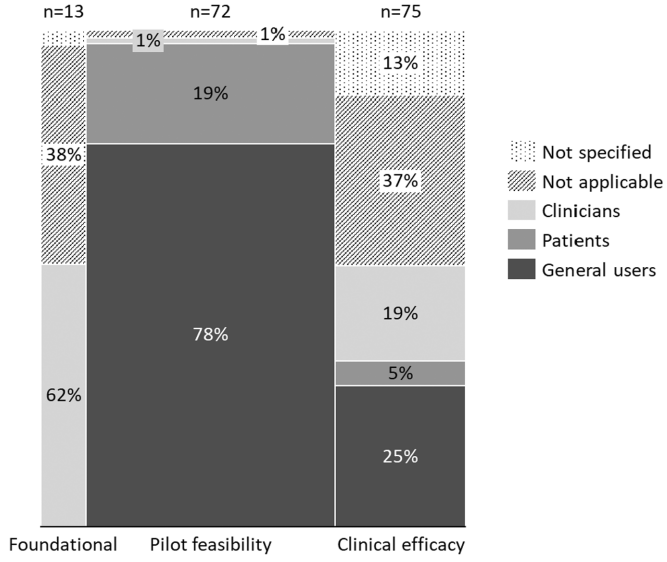

Figure 5 Evaluation participant types across study methodologies. Percentages indicate the distribution of participant types within each methodology. Subtotals and percentages are rounded to the nearest whole number, which may result in apparent summation discrepancies.

DISCUSSION

The rapid adoption of generative LLMs (e.g., *GPT-4*) reflects broader AI trends, but introduces unique risks in mental health contexts[185]. The complexity of mental health conditions, the subjective nature of diagnosis, and the need for contextual understanding further complicate AI integration[186-188]. While early rule-based systems such as *Woebot* prioritized safety through scripted dialogues, LLMs such as *Replika* now risk generating unvalidated advice due to their reliance on uncurated datasets[189]. This tension between innovation and safety, reflected in a December 2024 complaint by the American Psychological Association to the US Federal Trade Commission accusing a generative AI chatbot of harming children[190], underscores the need for structured validation frameworks and research to fill the gaps identified in our results.

A persistent challenge in the field is the misalignment between the marketed rhetoric of "AI-driven" systems and their underlying technical realities. Platforms such as *Woebot* and *Replika* market themselves as "AI-driven", yet the term "AI" remains ambiguously defined and is often employed without clear specification of the

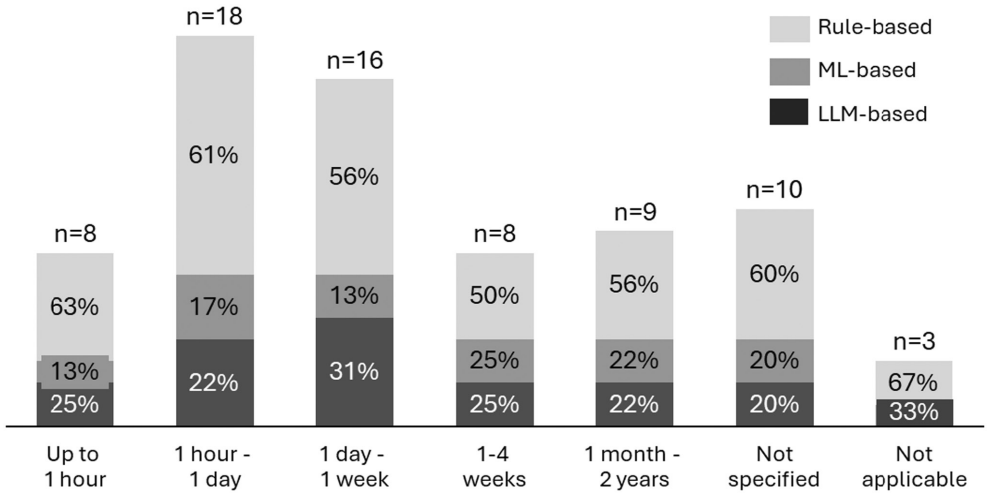

Figure 6 Distribution of study durations for the various chatbot architectures in the 72 pilot feasibility studies. Percentages indicate the proportion of chatbot architectures within each study duration category. Subtotals and percentages are rounded to the nearest whole number, which may result in apparent summation discrepancies. ML – machine learning, LLM – large language model.

underlying model architecture[191].

Early iterations of "AI" chatbots predominantly operated through scripted, rule-based interactions with only rudimentary machine learning enhancements. These rule-based tools, emblematic of the good old-fashioned artificial intelligence (GOFAI) paradigm, have been critiqued for their limited adaptability and depth, standing in sharp contrast to modern data-driven, connectionist approaches[192]. As both technological capabilities and stakeholder expectations evolve, clinicians and patients now increasingly expect AI to represent more sophisticated, dynamic and autonomous connectionist systems, such as LLMs, capable of generating contextually rich, free-form dialogue[193].

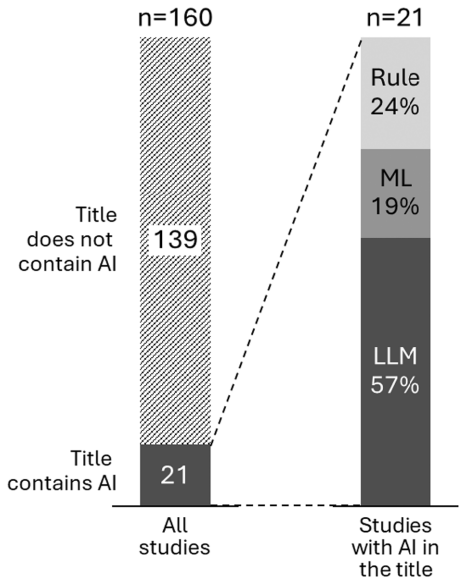

Figure 7 Usage of "AI" in study titles versus actual chatbot architectures. The left bar shows the proportion of all studies containing "AI" in the title. The right bar shows the percentage distribution of underlying chatbot architectures for the subset of studies with "AI" in the title. AI – artificial intelligence, Rule – rule-based, ML – machine learning, LLM – large language model.

This evolving expectation creates confusion, as legacy systems continue to be marketed under the same broad AI umbrella, making it difficult to distinguish between LLM-driven innovations and older rule-based or hybrid approaches. To address this, our classification system offers a structured way to categorize mental health chatbots based on architecture and function, distinguishing rule-based systems that operate through deterministic scripts, machine learning-based systems that enhance adaptability through data-driven models, and LLM-based systems that generate free-form, contextually rich dialogue. This architectural classification is critical, as it allows clinicians, researchers and regulators to appropriately evaluate chatbot capabilities, ensuring that expectations align with actual functionalities rather than misleading claims.

With the increasing number of chatbot studies and the sharp rise in LLM-based chatbot research, the need for standardized evaluation frameworks is more urgent than ever. While chatbots were once predominantly rule-based, requiring only basic assessments of functionality and engagement, the integration of machine learning- and LLM-based systems has generated complexity in evaluation. Unlike rule-based chatbots, which can be tested through predefined workflows, LLM chatbots introduce elements of unpredictability, requiring assessments beyond technical performance. Without a standardized nomenclature for evaluation, studies report outcomes inconsistently, making it difficult to compare findings across research. The introduction of a standardized three-tier evaluation continuum – foundational bench testing, pilot feasibility testing, and clinical efficacy testing – addresses this need by providing a clear progression of evidentiary rigor. It also fits well with recent calls for graded regulation of LLM systems, with certification linked to the role of the chatbot and rigor of its real-world efficacy[9].

Architectural analysis reveals that the focus and stage of evaluation vary significantly depending on chatbot type. Rule-based systems, lending themselves to structured interactions such as

symptom monitoring or delivering psychoeducational content, continue to be the most common architecture evaluated in clinical efficacy trials (65% of T3 studies). This likely reflects their longer history and suitability for interventions where predictability and safety are paramount. Machine learning-based chatbots, offering more adaptability than rule-based systems but lacking the generative fluency of LLMs, are represented modestly across all evaluation tiers (15-19%). In stark contrast, LLM-based systems heavily dominate foundational bench testing (77% of T1 studies), indicating that current research primarily investigates their technical capabilities, such as conversational quality or adherence to specific prompts, often in simulated scenarios. Despite their potential for nuanced, high-context interactions relevant to applications such as assessment or emotional support, LLMs are infrequently evaluated in T3 trials (16%). This suggests that, while LLMs are being actively explored for their technical promise, they have yet to undergo widespread, rigorous testing for clinical benefit in high-stakes mental health contexts, leaving a critical gap in evidence.

A major challenge highlighted by the T1-T3 framework is the heterogeneity in the types of evidence generated across the evaluation pipeline, often tied to the predominant chatbot architecture at each stage. Foundational bench testing (T1), where LLM-based studies are most prevalent, typically yields evidence related to technical performance – such as conversational coherence, linguistic accuracy, or safety in controlled tests. As studies progress to T2 pilot feasibility testing, the focus shifts towards usability, engagement, and user acceptance, evaluated across diverse groups including general users and patients. Rule-based studies are more common here (58%) than LLM-based studies (24%). Evidence of clinically meaningful impact, such as symptom reduction measured by validated scales over time, is primarily generated in T3 clinical efficacy testing, and rule-based systems are the main architecture assessed at this highest tier (65%).

The current concentration of LLM research in T1 and T2 stages means that these advanced models are often validated based on technical feasibility or short-term user experience metrics, rather than demonstrated therapeutic effectiveness. This disparity underscores a crucial limitation: strong performance in T1 stage or positive user feedback in T2 stage does not necessarily translate to T3 clinical efficacy. Furthermore, while T2 studies often specify diverse participant groups, T3 evaluations show less consistency in reporting participant characteristics, sometimes hindering the assessment of real-world applicability and long-term impact.

Ethical, safety and regulatory concerns are becoming increasingly critical as chatbots move closer to clinical deployment. LLM-based systems introduce considerable risks, including potentially greater data privacy violations than rule-based or machine learning-driven systems, algorithmic bias, and the potential for "hallucinations" (i.e., false or misleading responses) that could lead to harmful advice. All this may be exacerbated by the richer and more in-depth conversational capabilities of LLMs, which could encourage users to disclose more sensitive information[193]. Unlike rule-based chatbots, which are constrained to predefined responses, LLMs rely on large, uncurated datasets, making them susceptible to misinformation. These risks are not hypothetical: real-world examples, such as *Replika*'s backlash for generating inappropriate responses[194], illustrate the consequences of insufficient safeguards in generative AI systems[195]. While rule-based chatbots mitigate some of these risks through structured outputs, they lack the adaptive empathy needed for sustained mental health support. Machine learning-based systems occupy an intermediary position, balancing adaptability with limited generative capacity but often struggling with transparency and interpretability.

Addressing these risks requires regulatory bodies to establish clear certification pathways for LLM-driven chatbots, ensuring that innovations in generative AI are balanced with accountability and user safety. However, given the widespread availability and increasing use of base models (e.g., *GPT*[8], *Gemini*[196], *Claude*[197], *Llama*[20], *DeepSeek*[198]) for mental health applications, this alone may not be sufficient. An urgent research avenue is to independently establish the safety and clinical utility of these models, ideally through rapid, automated and repeatable evaluations as their capabilities continue to evolve.

Moving forward, mental health chatbot research must prioritize rigorous clinical efficacy trials for LLM-based systems, ensuring that chatbots progress beyond foundational and feasibility testing to real-world clinical validation. The development of standardized clinical endpoints, transparency in chatbot architectures, and regulatory alignment with AI-driven mental health tools will be essential in bridging the gap between feasibility and efficacy. As chatbots continue to evolve, robust validation methodologies will be necessary to ensure that they serve as effective, ethical, and clinically reliable tools for global mental health care.

CONCLUSIONS

Mental health chatbots have rapidly evolved from deterministic rule-based systems to sophisticated LLMs, signalling a transformative shift in digital psychiatry. Despite this promising advancement, our systematic analysis highlights a fragmented landscape with limited rigorous clinical validation, particularly concerning generative AI technologies. The proposed three-tier classification system clarifies evaluation rigor and reveals that most LLM-based interventions remain in early development phases.

Future research should prioritize rigorous clinical efficacy trials, transparent reporting of chatbot architecture, and ethical evaluations, to ensure that these technologies reliably enhance mental health care. Clinicians and policy makers must distinguish between marketing claims and technical realities, advocating for evidence-based standards analogous to established medical AI certification processes.

ACKNOWLEDGEMENTS

J. Torous is supported by the Argosy Foundation and Shifting Gears Foundation. Y. Hua and S. Siddals are joint first authors of this paper. Supplementary information on this study is available at https://osf.io/tcwv8/wiki/home.

REFERENCES

1. James SL, Abate D, Abate KH et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195

countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018;392:1789-858.
2. World Health Organization. Depression and other common mental disorders. www.who.int.
3. World Health Organization. Mental health. www.who.int.
4. World Health Organization. Global health workforce statistics database. www.who.int.
5. Patel V, Saxena S, Lund C et al. The Lancet Commission on global mental health and sustainable development. Lancet 2018;392:1553-98.
6. Abd-Alrazaq AA, Rababeh A, Alajlani M et al. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. J Med Internet Res 2020;22:e16021.
7. Brown TB, Mann B, Ryder N et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020:1877-901.
8. Ayers JW, Poliak A, Dredze M et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183:589-96.
9. Rajpurkar P, Topol EJ. A clinical certification pathway for generalist medical AI systems. Lancet 2025;405:20.
10. Weizenbaum J. ELIZA – a computer program for the study of natural language communication between man and machine. Communications of the ACM 1966;9:36-45.
11. Darcy A. Why generative AI is not yet ready for mental healthcare. Woebot Health, March 1, 2023.
12. Parmar P, Ryu J, Pandya S et al. Health-focused conversational agents in person-centered care: a review of apps. NPJ Digit Med 2022;5:1-9.
13. Yeh PL, Kuo WC, Tseng BL et al. Does the AI-driven chatbot work? Effectiveness of the Woebot app in reducing anxiety and depression in group counseling courses and student acceptance of technological aids. Curr Psychol 2025;44:8133-45.
14. Bishop CM. Pattern recognition and machine learning. New York: Springer, 2006.
15. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
16. Mikolov T, Sutskever I, Chen K et al. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst 2013;26:3111-9.
17. Fitzpatrick KK, Darcy A, Vierhile M. 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 Ment Health 2017;4:e19.
18. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. JMIR mHealth uHealth 2018;6:e12106.
19. Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30:5998-6008.
20. Touvron H, Lavril T, Izacard G et al. LLaMA: open and efficient foundation language models. arXiv 2023;230213971.
21. Hatch SG, Goodman ZT, Vowels L et al. When ELIZA meets therapists: a Turing test for the heart and mind. PLoS Ment Health 2025;2:e0000145.
22. Na H, Hua Y, Wang Z et al. A survey of large language models in psychotherapy: current landscape and future directions. arXiv 2025;2502.11095.
23. Bender EM, McMillan-Major A, Shmitchell S et al. On the dangers of stochastic parrots: can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021:610-23.
24. Page MJ, McKenzie JE, Bossuyt PM et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:71.
25. Gabrielli S, Rizzi S, Carbone S et al. A chatbot-based coaching intervention for adolescents to promote life skills: pilot study. JMIR Hum Factors 2020;7:e16762.
26. Denecke K, Vaaheesan S, Arulnathan A. A mental health chatbot for regulating emotions (SERMO) – concept and usability test. IEEE Trans Emerg Topics Comput 2020;9:1170-82.
27. Hauser-Ulrich S, Künzli H, Meier-Peterhans D et al. A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): pilot randomized controlled trial. JMIR mHealth uHealth 2020;8:e15806.
28. Linden B, Tam-Seto L, Stuart H. Adherence of the Here4U App - Military Version to criteria for the development of rigorous mental health apps. JMIR Form Res 2020;4:e18890.
29. Gaffney H, Mansell W, Tai S. Agents of change: understanding the therapeutic processes associated with the helpfulness of therapy for mental health problems with relational agent MYLO. Digit Health 2020;6:2055207620911580.
30. Dosovitsky G, Pineda BS, Jacobson NC et al. Artificial intelligence chatbot for depression: descriptive study of usage. JMIR Form Res 2020;4:e17065.
31. Kawasaki M, Yamashita N, Lee YC et al. Assessing users' mental status from their journaling behavior through chatbots. Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents, 2020:32.
32. Lee YC, Yamashita N, Huang Y. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. Proceedings of the ACM Conference on Human-Computer Interaction, 2020;4:31.
33. de Gennaro M, Krumhuber EG, Lucas G. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. Front Psychol 2020;10:3061.
34. De Nieva JO, Joaquin JA, Tan CB et al. Investigating students' use of a mental health chatbot to alleviate academic stress. Proceedings of the 6th International ACM In-Cooperation HCI and UX Conference, 2020:1-10.
35. Daley K, Hungerbuehler I, Cavanagh K et al. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. Front Digit Health 2020;2:576361.
36. Narain J, Quach T, Davey M et al. Promoting wellbeing with Sunny, a chatbot that facilitates positive messages within social groups. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020:1-8.
37. Ryu H, Kim S, Kim D et al. Simple and steady interactions win the healthy mentality: designing a chatbot service for the elderly. Proceedings of the ACM Conference on Human-Computer Interaction, 2020;4:152.
38. Bennion MR, Hardy GE, Moore RK et al. Usability, acceptability, and effectiveness of web-based conversational agents to facilitate problem solving in older adults: controlled study. J Med Internet Res 2020;22:e16794.
39. Park S, Thieme A, Han J et al. "I wrote as if I were telling a story to someone I knew": designing chatbot interactions for expressive writing in mental health. Proceeding of the ACM Designing Interactive Systems Conference, 2021:926-41.
40. Williams R, Hopkins S, Frampton C et al. 21-Day Stress Detox: open trial of a universal well-being chatbot for young adults. Soc Sci 2021;10:416.
41. Chung K, Cho HY, Park JY. A chatbot for perinatal women's and partners' obstetric and mental health care: development and usability evaluation study. JMIR Med Inform 2021;9:e18607.
42. Loveys K, Sagar M, Pickering I et al. A digital human for delivering a remote loneliness and stress intervention to at-risk younger and older adults during the COVID-19 pandemic: randomized pilot trial. JMIR Ment Health 2021;8:e31586.
43. Prochaska JJ, Vogel EA, Chieng A et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. J Med Internet Res 2021;23:e24850.
44. Oliveira ALS, Matos LN, Junior MC et al. An initial assessment of a chatbot for rumination-focused cognitive behavioral therapy (RFCBT) in college students. Proceeding of the International Conference on Computational Science and Its Applications, 2021:549-64.
45. Klos MC, Escoredo M, Joerin A et al. Artificial intelligence-based chatbot for anxiety and depression in university students: pilot randomized controlled trial. JMIR Form Res 2021;5:e20678.
46. Ellis-Brush K. Augmenting coaching practice through digital methods. Int J Evid Based Coach Mentor 2021;Spec. Issue 15:187-97.
47. Dosovitsky G, Bunge EL. Bonding with Bot: user feedback of a chatbot for social isolation. Front Digit Health 2021;3:735053.
48. Valtolina S, Hu L. Charlie: a chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness. Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter, 2021:19.
49. Hungerbuehler I, Daley K, Cavanagh K et al. Chatbot-based assessment of employees' mental health: design process and pilot implementation. JMIR Form Res 2021;5:e21678.
50. Lavelle J, Dunne N, Mulcahy HE et al. Chatbot-delivered cognitive defusion versus cognitive restructuring for negative self-referential thoughts: a pilot study. Psychol Rec 2021;72:247-61.
51. Kaywan P, Ahmed K, Miao Y et al. DEPRA: an early depression detection analysis chatbot. Proceedings of the International Conference on Health Information Science, 2021:193-204.
52. Beilharz F, Sukunesan S, Rossell SL et al. Development of a positive body image chatbot (KIT) with young people and parents/carers: qualitative focus group study. J Med Internet Res 2021;23:e27807.
53. Narynov S, Zhumanov Z, Gumar A et al. Development of chatbot psychologist applying natural language understanding techniques. Proceedings of the 21st International Conference on Control, Automation and Systems, 2021:636-41.
54. Romanovskyi O, Pidbutska N, Knysh A. Elomia chatbot: the effectiveness of artificial intelligence in the fight for mental health. Proceedings of the 5th In-

ternational Conference on Computational Linguistics and Intelligent Systems, 2021:1215-24.
55. van Cuylenburg HC, Ginige T. Emotion Guru: a smart emotion tracking application with AI conversational agent for exploring and preventing depression. Proceedings of the International Conference on UK-China Emerging Technologies, 2021:1-6.
56. Mariamo A, Temcheff CE, Léger PM et al. Emotional reactions and likelihood of response to questions designed for a mental health chatbot among adolescents: experimental study. JMIR Hum Factors 2021;8:e24343.
57. Gabrielli S, Rizzi S, Bassi G et al. Engagement and effectiveness of a healthy-coping intervention via chatbot for university students during the COVID-19 pandemic: mixed methods proof-of-concept study. JMIR mHealth uHealth 2021;9:e27965.
58. Darcy A, Daniels J, Salinger D et al. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. JMIR Form Res 2021;5:e27868.
59. Bendig E, Erb B, Meißner E et al. Feasibility of a software agent providing a brief Intervention for Self-help to Uplift psychological wellbeing ("SISU"). A single-group pretest-posttest trial investigating the potential of SISU to act as therapeutic agent. Internet Interv 2021;24:100377.
60. Sia DE, Yu MJ, Daliva JL et al. Investigating the acceptability and perceived effectiveness of a chatbot in helping students assess their well-being. Proceedings of the Asian CHI Symposium, 2021:34-40.
61. Rakib AB, Rumky EA, Ashraf AJ et al. Mental healthcare chatbot using sequence-to-sequence learning and BiLSTM. Brain Informatics Lecture Notes in Computer Science, 2021:378-87.
62. Jang S, Kim JJ, Kim SJ et al. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: a development and feasibility/usability study. Int J Med Inform 2021;150:104440.
63. Dosovitsky G, Kim E, Bunge EL. Psychometric properties of a chatbot version of the PHQ-9 for adults and older adults. Front Digit Health 2021;3:645805.
64. Lee J, Liang B, Fong H. Restatement and question generation for counsellor chatbot. Proceedings of the 1st Workshop on NLP for Positive Impact, 2021:1-7.
65. Salhi I, Guemmat KE, Qbadou M et al. Towards developing a pocket therapist: an intelligent adaptive psychological support chatbot against mental health disorders in a pandemic situation. Indones J Electr Eng Comput Sci 2021;23:1200.
66. Kraus M, Seldschopf P, Minker W. Towards the development of a trustworthy chatbot for mental health applications. Proceedings of the 27th International Conference on MultiMedia Modeling, 2021:354-66.
67. Goonesekera Y, Donkin L. A cognitive behavioral therapy chatbot (Otis) for health anxiety management: mixed methods pilot study. JMIR Form Res 2022;6:e37877.
68. Leo AJ, Schuelke MJ, Hunt DM et al. A digital mental health intervention in an orthopedic setting for patients with symptoms of depression and/or anxiety: feasibility prospective cohort study. JMIR Form Res 2022;6:e34889.
69. Beredo JL, Ong EC. A hybrid response generation model for an empathetic conversational agent. International Conference on Asian Language Processing, 2022:300-5.
70. Rathnayaka P, Mills N, Burnett D et al. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. Sensors 2022;22:3653.
71. Vertsberger D, Naor N, Winsberg M. Adolescents' well-being while using a mobile artificial intelligence-powered acceptance commitment therapy tool: evidence from a longitudinal study. JMIR AI 2022;1:e38171.
72. Liu I, Xiao Y, Liu F et al. Assessing the effectiveness of using chatbots for positive psychological intervention: a randomized control study. Proceedings of the 10th International Symposium of Chinese CHI, 2022:227-34.
73. Medeiros L, Bosse T, Gerritsen C. Can a chatbot comfort humans? Studying the impact of a supportive chatbot on users' self perceived stress. IEEE Trans Hum Mach Syst 2022;52:343-53.
74. Nicol G, Wang R, Graham S et al. Chatbot-delivered cognitive behavioral therapy in adolescents with depression and anxiety during the COVID-19 pandemic: feasibility and acceptability study. JMIR Form Res 2022;6:e40242.
75. Nie J, Shao H, Zhao M et al. Conversational AI therapist for daily function screening in home environments. Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications, 2022:31-6.
76. Collins C, Arbour S, Beals N et al. Covid Connect: chat-driven anonymous story-sharing for peer support. Proceedings of the ACM Designing Interactive Systems Conference, 2022:301-18.
77. Harshini Raji VP, Maheswari PU. COVID-19 lockdown in India: an experimental study on promoting mental wellness using a chatbot during the coronavirus. Int J Ment Health Prom 2022;24:189-205.
78. Maeng W, Lee J. Designing and evaluating a chatbot for survivors of image-based sexual abuse. Proceedings of the CHI Conference on Human Factors in Computing, 2022:344.
79. Shah J, DePietro B, D'Adamo L et al. Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening. Int J Eat Disord 2022;55:1229-44.
80. Trappey AJC, Lin APC, Hsu KYK et al. Development of an empathy-centric counseling chatbot system capable of sentimental dialogue analysis. Processes 2022;10:930.
81. Leo AJ, Schuelke MJ, Hunt DM et al. Digital mental health intervention plus usual care compared with usual care only and usual care plus in-person psychological counseling for orthopedic patients with symptoms of depression or anxiety: cohort study. JMIR Form Res 2022;6:e36203.
82. Fitzsimmons-Craft EE, Chan WW, Smith et al. Effectiveness of a chatbot for eating disorders prevention: a randomized clinical trial. Int J Eat Disord 2022;55:343-53.
83. Liu I, Chen W, Ge Q et al. Enhancing psychological resilience with chatbot-based cognitive behavior therapy: a randomized control pilot study. Proceedings of the 10th International Symposium of Chinese CHI, 2022:216-21.
84. Beatty C, Malik T, Meheli S et al. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. Front Digit Health 2022;4.
85. He Y, Yang L, Zhu X et al. Mental health chatbot for young adults with depression symptoms during the COVID-19 pandemic: a single-blind, three-arm, randomized controlled trial. J Med Internet Res 2022;24:e40719.
86. Santa-Cruz J, Moran L, Tovar M et al. Mobilizing digital technology to implement a population-based psychological support response during the COVID-19 pandemic in Lima, Peru. Glob Ment Health 2022;9:355-65.
87. Boyd K, Potts C, Bond R et al. Usability testing and trust analysis of a mental health and wellbeing chatbot. Proceedings of the 33rd European Conference on Cognitive Ergonomics, 2022:18.
88. Burger F, Neerincx MA, Brinkman WP. Using a conversational agent for thought recording as a cognitive therapy task: feasibility, content, and feedback. Front Digit Health 2022;4.
89. Liu H, Peng H, Song X et al. Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. Internet Interv 2022;27:100495.
90. Schick A, Feine J, Morana S et al. Validity of chatbot use for mental health assessment: experimental study. JMIR mHealth uHealth 2022:e28082.
91. Todorov M, Avramova-Todorova G, Dimitrova et al. Virtual assisted technologies as a helping tool for therapists in assessment of anxiety. Outcomes of a pilot trial with chatbot assistance. Proceedings of the International Symposium on Bioinformatics and Biomedicine, 2022:60-6.
92. Sanabria G, Greene KY, Tran JT et al. "A great way to start the conversation": evidence for the use of an adolescent mental health chatbot navigator for youth at risk of HIV and other STIs. J Technol Behav Sci 2023;8:382-91.
93. Chung LL, Kang J. "I'm hurt too": the effect of a chatbot's reciprocal self-disclosures on users' painful experiences. Arch Design Res 2023;36:67-85.
94. Sabour S, Zhang W, Xiao X et al. A chatbot for mental health support: exploring the impact of Emohaa on reducing mental distress in China. Front Digit Health 2023;5.
95. Escobar-Viera CG, Porta G, Coulter RW et al. A chatbot-delivered intervention for optimizing social media use and reducing perceived isolation among rural-living LGBTQ+ youth: development, acceptability, usability, satisfaction, and utility. Internet Interv 2023;34:100668.
96. Brinsley J, Singh B, Maher CA. A digital lifestyle program for psychological distress, wellbeing and return-to-work: a proof-of-concept study. Arch Phys Med Rehabil 2023;104:1903-12.
97. Potts C, Lindstrm F, Bond R et al. A multilingual digital mental health and well-being chatbot (ChatPal): pre-post multicenter intervention study. J Med Internet Res 2023;25:e43051.
98. Alonso-Mencía J, Castro-Rodríguez M, Herrero-Pinilla B et al. ADELA: a conversational virtual assistant to prevent delirium in hospitalized older persons. J Supercomput 2023;79:17670-90.
99. Gabor-Siatkowska K, Sowanski M, Rzatkiewicz R et al. AI to train AI: using ChatGPT to improve the accuracy of a therapeutic dialogue system. Electronics 2023;12:4694.
100. Wrightson-Hester AR, Anderson G, Dunstan J et al. An artificial therapist (Manage Your Life Online) to support the mental health of youth: co-design and case series. JMIR Hum Factors 2023;10:e46849.
101. Omarov B, Zhumanov Z, Gumar A et al. Artificial intelligence enabled mobile chatbot psychologist using AIML and cognitive behavioral therapy. Int J Adv

Comput Sci Appl 2023;14.
102. Viduani A, Cosenza V, Fisher HL et al. Assessing mood with the identifying depression early in adolescence chatbot (IDEABot): development and implementation study. JMIR Hum Factors 2023;10:e44388.
103. Ollier J, Suryapalli P, Fleisch E et al. Can digital health researchers make a difference during the pandemic? Results of the single-arm, chatbot-led Elena+: care for COVID-19 interventional study. Front Public Health 2023;11:1185702.
104. Yorita A, Egerton S, Chan C et al. Chatbot persona selection methodology for emotional support. Proceedings of the 62nd Annual Conference of the Society of Instrument and Control Engineers, 2023:333-8.
105. Lee CH, Liaw GH, Yang WC et al. Chatbot-assisted therapy for patients with methamphetamine use disorder. Front Psychiatry 2023;14:1159399.
106. Selaskowski B, Reiland M, Schulze M et al. Chatbot-supported psychoeducation in adult attention-deficit hyperactivity disorder: randomised controlled trial. BJPsych Open 2023;9:e192.
107. Sezgin E, Chekeni F, Lee J et al. Clinical accuracy of large language models and Google search responses to postpartum depression questions: cross-sectional study. J Med Internet Res 2023;25:e49240.
108. Kang E, Kang YA. Counseling chatbot design: the effect of anthropomorphic chatbot characteristics on user self-disclosure and companionship. Int J Hum-Comput Interact 2023;40:2781-95.
109. Lee M, Contreras Alejandro J, IJsselsteijn W. Cultivating gratitude with a chatbot. Int J Hum-Comput Interact 2023;40:4957-72.
110. Islam A, Chaudhry BM. Design validation of a relational agent by COVID-19 patients: mixed methods study. JMIR Hum Factors 2023;10:e42740.
111. Dosovitsky G, Bunge E. Development of a chatbot for depression: adolescent perceptions and recommendations. Child Adolesc Ment Health 2023;28:124-7.
112. Park G, Chung J, Lee S. Effect of AI chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: a serial mediation model. Curr Psychol 2023;42:28663-73.
113. Mishra K, Priya P, Burja M et al. e-THERAPIST: I suggest you to cultivate a mindset of positivity and nurture uplifting thoughts. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023:13952-67.
114. Iglesias M, Sinha C, Vempati R et al. Evaluating a digital mental health intervention (Wysa) for workers' compensation claimants: pilot feasibility study. J Occup Environ Med 2023;65:e93-9.
115. Kang A, Hetrick S, Cargo T et al. Exploring young adults views about Aroha, a chatbot for stress associated with the COVID-19 pandemic: interview study among students. JMIR Form Res 2023;7:e44556.
116. Suharwardy S, Ramachandran M, Leonard SA et al. Feasibility and impact of a mental health chatbot on postpartum mental health: a randomized controlled trial. AJOG Glob Rep 2023;3:100165.
117. Bird JJ, Lotfi A. Generative transformer chatbots for mental health support: a study on depression and anxiety. Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments, 2023:475-9.
118. Sharma A, Lin IW, Miner AS et al. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. Nat Mach Intell 2023;5:46-57.
119. Han HJ, Mendu S, Jaworski BK et al. Preliminary evaluation of a conversational agent to support self-management of individuals living with posttraumatic stress disorder: interview study with clinical experts. JMIR Form Res 2023;7:e45894.
120. Rodriguez-Martinez A, Amezcua-Aguilar T, Cortés-Moreno J et al. Qualitative analysis of conversational chatbots to alleviate loneliness in older adults as a strategy for emotional health. Healthcare 2023;12:62.
121. Heston TF. Safety of large language models in addressing depression. Cureus 2023;15:e50729.
122. Mancone S, Diotaiuti P, Valente G et al. The use of voice assistant for psychological assessment elicits empathy and engagement while maintaining good psychometric properties. Behav Sci 2023;13:550.
123. Sinha C, Meheli S, Kadaba M. Understanding digital mental health needs and usage with an artificial intelligence-led mental health app (Wysa) during the COVID-19 pandemic: retrospective analysis. JMIR Form Res 2023;7:e41913.
124. Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of LLM-based agents for mental well-being. AMIA Annu Symp Proc 2023;2023:1105-14.
125. Inkster B, Kadaba M, Subramanian V. Understanding the impact of an AI-enabled conversational agent mobile app on users' mental health and well-being with a self-reported maternal event: a mixed method real-world data mHealth study. Front Glob Womens Health 2023;4:1084302.
126. Matheson EL, Smith HG, Amaral ACS et al. Using chatbot technology to improve Brazilian adolescents' body image and mental health at scale: randomized controlled trial. JMIR mHealth uHealth 2023;11:e39934.
127. Anmella G, Sanabra M, Primé-Tous M et al. Vickybot, a chatbot for anxiety-depressive symptoms and work-related burnout in primary care and health care professionals: development, feasibility, and potential effectiveness studies. J Med Internet Res 2023;25:e43293.
128. Ge Q, Liu L, Zhang H et al. Designing Philobot: a chatbot for mental health support with CBT techniques. Proceedings of the Chinese Intelligent Automation Conference, 2023:361-71.
129. Siddals S, Torous J, Coxon A. "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. Npj Ment Health Res 2024;3:48.
130. da Costa FL, Matzenbacher LS, Maia IS et al. "She has become my best friend": a qualitative study on the perspective of elderly with type 2 diabetes regarding the use of an interactive virtual assistant device for diabetes care and mental health promotion. Acta Diabetol 2024; doi: 10.1007/s00592-024-02377-z.
131. Mancinelli E, Magnolini S, Gabrielli S et al. A chatbot (Juno) prototype to deploy a behavioral activation intervention to pregnant women: qualitative evaluation using a multiple case study. JMIR Form Res 2024;8:e58653.
132. Yasukawa S, Tanaka T, Yamane K et al. A chatbot to improve adherence to internet-based cognitive-behavioural therapy among workers with subthreshold depression: a randomised controlled trial. BMJ Ment Health 2024;27:e300881.
133. Ulrich S, Lienhard N, Künzli H et al. A chatbot-delivered stress management coaching for students (MISHA app): pilot randomized controlled trial. JMIR mHealth uHealth 2024;12:e54945.
134. Boian R, Bucur AM, Todea D et al. A conversational agent framework for mental health screening: design, implementation, and usability. Behav Inform Technol 2024; doi: 10.1080/0144929X.2024.2332934.
135. Allan DD. A mobile messaging-based conversational agent-led stress mindset intervention for New Zealand small-to-medium-sized enterprise owner-managers: effectiveness and acceptability study. Behav Inf Technol 2024;43:2785-98.
136. Shidara K, Tanaka H, Adachi H et al. Adapting the number of questions based on detected psychological distress for cognitive behavioral therapy with an embodied conversational agent: comparative study. JMIR Form Res 2024;8:e50056.
137. So J-H, Chang J, Kim E et al. Aligning large language models for enhancing psychiatric interviews through symptom delineation and summarization: pilot study. JMIR Form Res 2024;8:e58418.
138. Foran HM, Kubb C, Mueller J et al. An automated conversational agent self-help program: randomized controlled trial. J Med Internet Res 2024;26:e53829.
139. Marmol-Romero AM, Garcia-Vega M, Garcia-Cumbreras M et al. An empathic GPT-based chatbot to talk about mental disorders with Spanish teenagers. Int J Hum-Comput Interact 2024; doi: 10.1080/10447318.2024.2344355.
140. Ruggiano N, Brown EL, Clarke PJ et al. An evidence-based IT program with chatbot to support caregiving and clinical care for people with dementia: the CareHeroes development and usability pilot. JMIR Aging 2024;7:e57308.
141. Han HJ, Mendu S, Jaworski BK et al. Assessing acceptance and feasibility of a conversational agent to support individuals living with post-traumatic stress disorder. Digit Health 2024;10:20552076241286133.
142. Maurya RK, Montesinos S, Bogomaz M et al. Assessing the use of ChatGPT as a psychoeducational tool for mental health practice. Couns Psychother Res 2024; doi: 10.1002/capr.12759.
143. Heissler R, Jonaa J, Carre N et al. Can AI digital personas for well-being provide social support? A mixed-method analysis of user reviews. Hum Behav Emerg Technol 2024;2024:6738001.
144. Hodson N, Williamson S. Can large language models replace therapists? Evaluating performance at simple cognitive behavioral therapy tasks. JMIR AI 2024;3:e52500.
145. Li Y, Chung TY, Lu W et al. Chatbot-based mindfulness-based stress reduction program for university students with depressive symptoms: intervention development and pilot evaluation. J Am Psychiatr Nurses Assoc 2024; doi: 10.1177/10783903241302092.
146. Striegl J, Fekih F, Weber G et al. Chatbot-based mood and activity journaling for resource-oriented CBT support of students. Proceedings of the International Conference on Human-Computer Interaction, 2024:177-88.
147. Dergaa I, Fekih-Romdhane F, Hallit S et al. ChatGPT is not ready yet for use in providing mental health assessment and interventions. Front Psychiatry 2024;14:1277756.
148. Gargari OK, Habibi G, Nilchian N et al. Comparative analysis of large language models in psychiatry and mental health: a focus on GPT, AYA, and Nemotron-3-8B. Asian J Psychiatr 2024;99:104148.
149. Liu T, Zhao H, Liu Y et al. ComPeer: a generative conversational agent for pro-

active peer support. Proceedings of the 37th ACM Symposium on User Interface Software and Technology, 2024:1-22.
150. Aghakhani S, Rousseau A, Mizrahi S et al. Conversational agent utilization patterns of individuals with autism spectrum disorder. J Technol Behav Sci 2024; doi: 10.1007/s41347-024-00451-5.
151. Ferrández A, Lavigne-Cerván R, Peral J et al. CuentosIE: can a chatbot about "tales with a message" help to teach emotional intelligence? PeerJ Comput Sci 2024;10:e1866.
152. Chiauzzi E, Williams A, Mariano TY et al. Demographic and clinical characteristics associated with anxiety and depressive symptom outcomes in users of a digital mental health intervention incorporating a relational agent. BMC Psychiatry 2024;24:79.
153. Raut R, Kolambe S, Borkar P et al. Depression therapy chat-bot using natural language processing. Int J Intell Syst Appl Eng 2024;12:181-7.
154. Nasiri Y, Fulda N. Designing a language-model-based chatbot that considers user's personality profile and emotions to support caregivers of people with dementia. Proceedings of the 1st International OpenKG Workshop: Large Knowledge-Enhanced Models, 2024:59-70.
155. Abilkaiyrkyzy A, Laamarti F, Hamdi M et al. Dialogue system for early mental illness detection: toward a digital twin solution. IEEE Access 2024;12:2007-24.
156. Kleinau EF, Lamba T, Jaskiewicz W et al. Effectiveness of a chatbot in improving the mental wellbeing of health workers in Malawi during the COVID-19 pandemic: a randomized, controlled trial. PLoS One 2024;19:e0303370.
157. MacNeill AL, Doucet S, Luke A. Effectiveness of a mental health chatbot for people with chronic diseases: randomized controlled trial. JMIR Form Res 2024;8:e50025.
158. Karkosz S, Szymanski R, Sanna K et al. Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: randomized controlled trial. JMIR Form Res 2024;8:e47960.
159. Chew HSJ, Chew NW, Loong SSE et al. Effectiveness of an artificial intelligence-assisted app for improving eating behaviors: mixed methods evaluation. J Med Internet Res 2024;26:e46036.
160. Vereschagin M, Wang AY, Richardson CG et al. Effectiveness of the Minder mobile mental health and substance use intervention for university students: randomized controlled trial. J Med Internet Res 2024;26:e54287.
161. Schillings C, Meißner E, Erb B et al. Effects of a chatbot-based intervention on stress and health-related parameters in a stressed sample: randomized controlled trial. JMIR Ment Health 2024;11:e50454.
162. Fitzsimmons-Craft EE, Rackoff GN, Shah J et al. Effects of chatbot components to facilitate mental health services use in individuals with eating disorders following online screening: an optimization randomized controlled trial. Int J Eat Disord 2024;57:2204-16.
163. Yeo YH, Peng Y, Mehra M et al. Evaluating for evidence of sociodemographic bias in conversational AI for mental health support. Cyberpsychol Behav Soc Netw 2024; doi: 10.1089/cyber.2024.0199.
164. Sinha C, Dinesh D, Phang YS et al. Examining a brief web and longitudinal app-based intervention [Wysa] for mental health support in Singapore during the COVID-19 pandemic: mixed-methods retrospective observational study. Front Digit Health 2024;6:1443598.
165. Bowman R, Cooney O, Newbold JW et al. Exploring how politeness impacts the user experience of chatbots for mental health support. Int J Hum Comput Stud 2024;184:103181.
166. Kim Y, Kang Y, Kim B et al. Exploring the role of engagement and adherence in chatbot-based cognitive training for older adults: memory function and mental health outcomes. Behav Inform Technol 2024; doi: 10.1080/0144929x.2024.2362406.
167. Sharma A, Rushton K, Lin IW et al. Facilitating self-guided mental health interventions through human-language model interaction: a case study of cognitive restructuring. Proceedings of the CHI Conference on Human Factors in Computing Systems, 2024:1-29.
168. Berrezueta-Guzman S, Kandil M, Martin-Ruiz ML et al. Future of ADHD care: evaluating the efficacy of ChatGPT in therapy enhancement. Healthcare 2024;12:683.
169. Hu Z, Hou H, Ni S. Grow with your AI buddy: designing an LLMs-based conversational agent for the measurement and cultivation of children's mental resilience. Proceedings of the 23rd Annual ACM Interaction Design and Children Conference, 2024:811-7.
170. Park G, Chung J, Lee S. Human vs. machine-like representation in chatbot mental health counseling: the serial mediation of psychological distance and trust on compliance intention. Curr Psychol 2024;43:4352-63.
171. Lee J, Lee J-G, Lee D. Influence of rapport and social presence with an AI psychotherapy chatbot on users' self-disclosure. Int J Hum-Comput Interact 2024;40:1620-31.
172. Liu I, Liu F, Xiao Y et al. Investigating the key success factors of chatbot-based positive psychology intervention with retrieval- and generative pre-trained transformer (GPT)-based chatbots. Int J Hum-Comput Interact 2024;41:341-52.
173. Dinesh DN, Rao MN, Sinha C. Language adaptations of mental health interventions: user interaction comparisons with an AI-enabled conversational agent (Wysa) in English and Spanish. Digit Health 2024; doi: 10.1177/20552076241255616.
174. Qi Y. Pilot quasi-experimental research on the effectiveness of the Woebot AI chatbot for reducing mild depression symptoms among athletes. Int J Hum-Comput Interact 2024;41:452-9.
175. Mori D, Matsumoto K, Kang X et al. SMACS: Stress Management AI Chat System. Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2024:167-74.
176. Vossen W, Szymanski M, Verbert K. The effect of personalizing a psychotherapy conversational agent on therapeutic bond and usage intentions. Proceedings of the 29th International Conference on Intelligent User Interfaces, 2024:761-71.
177. Zsoldos I, Tran E, Fournier H et al. The value of a virtual assistant to improve engagement in computerized cognitive training at home: exploratory study. JMIR Rehabil Assist Technol 2024;11:e48129.
178. Andrade-Arenas L, Yactayo-Arias C, Pucuhuayla-Revatta F. Therapy and emotional support through a chatbot. Int J Online Biomed Eng 2024;20:114-30.
179. Thunstrm AO, Carlsen HK, Ali L et al. Usability comparison among healthy participants of an anthropomorphic digital human and a text-based chatbot as a responder to questions on mental health: randomized controlled trial. JMIR Hum Factors 2024;11:e54581.
180. Farrand P, Raue PJ, Ward E et al. Use and engagement with low-intensity cognitive behavioral therapy techniques used within an app to support worry management: content analysis of log data. JMIR mHealth uHealth 2024;12:e47321.
181. Chaudhry BM, Debi HR. User perceptions and experiences of an AI-driven conversational agent for mental health support. Mhealth 2024;10:22-2.
182. Chou YH, Lin C, Lee SH et al. User-friendly chatbot to mitigate the psychological stress of older adults during the COVID-19 pandemic: development and usability study. JMIR Form Res 2024;8:e49462.
183. Deepaisarn S, Imkome E, Wongpatikaseree K et al. Validation of a Thai artificial chatmate designed for cheering up the elderly during the COVID-19 pandemic. F1000Research 2024;11:1411.
184. Kostenius C, Lindstrom F, Potts C et al. Young peoples' reflections about using a chatbot to promote their mental wellbeing in northern periphery areas – a qualitative study. Int J Circumpolar Health 2024;83:2369349.
185. Li H, Zhang R, Lee YC et al. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. NPJ Digit Med 2023;6:236.
186. Torous J, Myrick K. Aguilera A. The need for a new generation of digital mental health tools to support more accessible, effective and equitable care. World Psychiatry 2023;22:1-2.
187. Torous J, Blease C. Generative artificial intelligence in mental health care: potential benefits and current challenges. World Psychiatry 2024;23:1-2.
188. Galderisi S, Appelbaum PS, Gill N et al. Ethical challenges in contemporary psychiatry: an overview and an appraisal of possible strategies and research needs. World Psychiatry 2024;23:364-86.
189. Lawrence HR, Schneider RA, Rubin SB et al. The opportunities and risks of large language models in mental health. JMIR Ment Health 2024;11:e59479.
190. Evans AC. Letter to the Federal Trade Commission regarding concerns about the perils and unintended consequences to the public resulting from the underregulated development and deceptive deployment of generative AI or enabled technologies. American Psychological Association, December 20, 2024.
191. Wang P. On defining artificial intelligence. J Artif Gen Intell 2019;10:1-37.
192. Boden MA. Gofai. In: Frankish K, Ramsey W (eds). The Cambridge handbook of artificial intelligence. Cambridge: Cambridge University Press, 2014:89-107.
193. Nong P, Ji M. Expectations of healthcare AI and the role of trust: understanding patient views on how AI will impact cost, access, and patient-provider relationships. J Am Med Inform Assoc 2025;32:795-9.
194. Roose K. Can A.I. be blamed for a teen's suicide? New York Times, October 23, 2024.
195. Blease C, Torous J. ChatGPT and mental healthcare: balancing benefits with risks of harms. BMJ Ment Health 2023;26:e300884.
196. Team G, Anil R, Borgeaud S et al. Gemini: a family of highly capable multimodal models. arXiv 2025;231211805.
197. Anthropic. Introducing the next generation of Claude. www.anthropic.com.
198. Deep Seek-AI, Liu A, Feng B et al. DeepSeek-V3 technical report. arXiv 2025;241219437.

DOI:10.1002/wps.21352