



International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	CHATBOT-DELIVERED COGNITIVE BEHAVIOURAL THERAPY: A STRUCTURED REVIEW OF EFFICACY, USER ENGAGEMENT, AND EQUITY OF ACCESS
DOI	https://doi.org/10.31435/ijitss.4(52).2026.6374
RECEIVED	16 July 2026
ACCEPTED	25 September 2026
PUBLISHED	05 October 2026
LICENSE	 The article is licensed under a Creative Commons Attribution 4.0 International License .

© The author(s) 2026.

This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

CHATBOT-DELIVERED COGNITIVE BEHAVIOURAL THERAPY: A STRUCTURED REVIEW OF EFFICACY, USER ENGAGEMENT, AND EQUITY OF ACCESS

Krzysztof Konopka (Corresponding Author, Email: normalemail9994@gmail.com)
Pomeranian Medical University, Szczecin, West Pomerania, Poland
ORCID ID: 0009-0003-4055-0074

Faustyna Magdalena Kołodziej
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0002-8778-7100

Alicja Kwiatkowska
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0003-1027-2114

Jakub Juskowiak
Poznań University of Medical Sciences, Poznań, Poland
ORCID ID: 0009-0001-0372-400X

Mateusz Lis
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0006-8441-8296

Zuzanna Julia Komin
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0001-6284-6410

Maryan Liasota
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0006-0266-4849

Eliza Zofia Kabat
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0003-6531-967X

Michalina Justyna Kozłowska
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0003-9561-2957

Maciej Kamil Majchrzak
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0007-4739-0092

ABSTRACT

Background: Chatbots are increasingly used to deliver cognitive behavioural therapy (CBT) via mobile and web applications. Many mental health systems cannot meet demand with face-to-face care alone. CBT chatbots have therefore been introduced in app form, but data on outcomes, dropout, and unequal uptake remain limited.

Objective: This review evaluates whether CBT chatbots reduce anxiety and depression symptoms, which user- and design-related factors are linked to engagement and dropout, and which structural barriers limit who can practically use these tools.

Methodology: Peer-reviewed studies from 2018 to 2026 were identified through searches in PubMed/MEDLINE and Google Scholar. Included studies examined automated CBT chatbots (e.g., Woebot, Wysa, Youper, Limbic Care) and reported clinical outcomes, engagement or dropout, and/or access-related findings. Twenty-four publications were synthesised thematically.

Results: Short-term trials and observational studies reported reductions in anxiety and depression symptoms, with small-to-moderate effects¹. Generative conversational designs were linked to higher engagement than static digital programmes in at least one RCT, which reported roughly 2.4 times the use frequency of a generative CBT app relative to workbook controls³. Attrition remained notable; one meta-analysis estimated intervention-arm dropout at 21.84%⁵. Unequal access persisted because of connectivity, device availability, and digital literacy, despite the privacy and round-the-clock availability of app-based care²³.

Conclusion: CBT chatbots may serve as useful low-intensity interventions, but their public health value depends on improved retention and fairer access. Implementation should address digital literacy, culturally appropriate design, and governance that does not restrict validated tools to already advantaged users.

KEYWORDS

Conversational Agents, Cognitive Behavioural Therapy, Digital Mental Health, Health Equity, User Engagement, Artificial Intelligence

CITATION

Konopka, K., Kołodziej, F. M., Kwiatkowska, A., Juskowiak, J., Lis, M., Komin, Z. J., Liasota, M., Kabat, E. Z., Kozłowska, M. J., & Majchrzak, M. K. (2026). Chatbot-delivered cognitive behavioural therapy: A structured review of efficacy, user engagement, and equity of access. *International Journal of Innovative Technologies in Social Science*, 4(52). [https://doi.org/10.31435/ijitss.4\(52\).2026.6374](https://doi.org/10.31435/ijitss.4(52).2026.6374)

COPYRIGHT

© The author(s) 2026. This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

1. Introduction

1.1 Background and Clinical Context

Many countries report too few mental health clinicians relative to patient demand. Service waiting times were already lengthy before COVID-19, and the pandemic period placed additional strain on outpatient care⁹. In many European health systems, stepped-care pathways are meant to offer brief psychological support before specialist referral, but that first step is still often under-resourced in routine practice⁶.

CBT remains a first-line treatment for anxiety and depression, but fees, travel, waiting lists, and stigma still prevent many people from starting face-to-face therapy²³. Long delays before the first appointment can allow symptoms to worsen, and some patients never attend after referral⁶. Where services do exist, clinician time is concentrated on the most severe cases, while large numbers of people with subthreshold or mild symptoms receive no structured care at all¹.

Internet-based CBT programmes and other digital tools were introduced to supplement clinic services. The earliest versions often used fixed readings or workbook tasks; controlled studies reported benefit, but routine uptake and homework completion were frequently low without ongoing contact with a therapist⁶. Message-based chatbot apps were later developed to walk users through CBT steps in an interactive format⁴. Their appeal lies partly in immediacy: users can open a session outside office hours without scheduling or travel.

National mental health plans now mention digital tools more often, but funding and staff training to use them in clinics remain patchy⁶. Chatbots sit between pure self-help apps and therapist-led iCBT, which is why commissioners look at them for scalable first-line support.

In the European Union, digital mental health appears in both national recovery plans and cross-border telehealth discussions, yet service-level adoption remains uneven. Chatbot CBT is often framed as a way to extend brief psychological support when psychiatrists and psychologists are scarce, especially in primary care and community settings.

1.2 The Rise of Chatbot-Delivered Mental Health Interventions

Unlike static modules, chatbots collect responses turn by turn in a chat screen. Established products such as Woebot, Wysa, and Youper teach techniques such as thought challenging, mood logging, and behavioural activation through short exchanges². Their backends typically pair clinician-written scripts with keyword matching or intent classification^{2, 12}. That structure reduces safety risks and keeps content close to protocol, but it also caps conversational flexibility.

A smaller set of newer apps, including Limbic Care, applies large language models to adapt homework and replies³. Published trials for these generative systems remain fewer than for rule-based tools, although early RCT evidence suggests they may increase session frequency and total time in app-based CBT³. Reviews of the wider chatbot literature note rapid product turnover and uneven reporting of implementation features^{21, 24}.

These products are not the same as general chatbots such as ChatGPT. Evaluated mental health bots use fixed clinical pathways, crisis routing in several commercial releases, and outcome measures tied to symptoms rather than chat quality alone^{12, 20}. Where wellness apps end and treatment tools begin is still contested in regulation⁷.

1.3 From Technical Performance to Equity and Engagement

Positive symptom scores in a trial mean little if users open the app only briefly or cannot install it on their phone¹. For that reason, retention, session frequency, and dropout need to be reported next to PHQ-9, GAD-7, and similar scales⁵. Engagement also affects dose: CBT techniques only work if users return and practise between sessions⁴.

Round-the-clock access and private use at home are frequently cited as advantages of mental health apps⁹. Published work also documents gaps in internet coverage, device ownership, digital skills, and training data that reflect narrow user groups²³. Evidence-based products with regulatory clearance may reach insured or technology-confident users, while others use unregulated chat tools⁷, which can leave existing inequalities unchanged or make them worse.

Similar equity concerns appear elsewhere in digital health: availability of technology does not mean equal uptake across income, region, or ethnic groups²². For chatbots, language, literacy, and trust in automated support add further barriers^{10, 23}. We therefore treat chatbots as interventions shaped by health systems, not isolated apps.

1.4 Research Gap and Rationale

Published work on CBT chatbots often treats clinical outcomes, user retention, and access as separate topics. Systematic reviews and meta-analyses have strengthened evidence on short-term symptom change^{1, 13}, and narrative reviews have catalogued implementation features^{12, 21}. Fewer papers explicitly link trial results for named platforms to evidence on who is excluded during real-world rollout^{5, 23}.

Most included trials were run by academic groups or with app developers. Independent follow-up of commercially updated versions is still limited, which matters for services that procure a product after the original publication²¹.

Earlier reviews vary in scope—some cover any mental health bot, others one architecture or diagnosis^{24, 12}. This review keeps the focus on CBT chatbots and groups evidence around symptom change, retention, and fair access.

1.5 Aim of the Review

This review addresses three questions: What do studies report on anxiety and depression outcomes after chatbot-delivered CBT? Which user and design factors are associated with continued use and dropout? Which structural barriers shape equitable access?

The synthesis combines RCTs, observational studies, and qualitative work through a socio-technical lens for clinicians, researchers, and policy staff weighing chatbot CBT against other care options.

We do not produce a league table of commercial products. The aim is to show what the literature reports on outcomes, engagement, and access so readers can judge relevance to their own setting.

From a social science perspective, the main question is not only whether chatbots work in trials but how they reshape access to care, labour in mental health services, and expectations of self-management among users who already face structural disadvantage²³.

Stakeholder perspectives also diverge. Developers emphasise scalability and engagement metrics; clinicians worry about risk, scope of practice, and whether automated tools dilute therapeutic relationships^{8, 14}. Patients may value anonymity and immediacy yet still want a human option when symptoms worsen¹⁹. A review aimed at social science and policy readers therefore needs to connect trial outcomes with these organisational and relational tensions, not only with effect sizes.

That framing matches the journal focus on how technology interacts with policy, service organisation, and social inequality rather than isolated software performance.

Healthcare planners increasingly ask whether chatbots belong in the lowest step of stepped-care ladders or only as adjuncts after initial human contact. Stepped-care models in the United Kingdom and Nordic countries already use digitally delivered CBT for mild anxiety and depression, but signposting rules differ between services^{6, 8}. In Poland and wider Central Europe, outpatient psychology capacity remains constrained relative to demand, which makes low-intensity digital options attractive on paper, yet reimbursement pathways for non-prescription apps are still unclear⁷. For chatbot CBT to function as more than a wellness download, local services need agreed criteria for when automated support is sufficient and when users must be escalated.

2. Methodology

2.1 Research Design

The review used a structured narrative design. Statistical pooling across trials was not performed; instead, RCTs, observational studies, and qualitative reports were reviewed within a socio-technical framework covering clinical outcomes, user behaviour, and access¹². This approach was chosen because trials used different controls, follow-up lengths, and engagement measures, so a single pooled estimate would need stronger homogeneity than the literature offers.

We used existing chatbot typologies^{21, 24} but kept the focus on CBT delivery. Results are presented by theme, noting where design choices and service setting may have influenced findings.

2.2 Data Source and Search Strategy

Peer-reviewed papers from January 2018 to March 2026 were identified in PubMed/MEDLINE. Google Scholar was used to locate additional interdisciplinary and recent publications, including human-computer interaction and digital health journals that may be indexed with delay.

In PubMed/MEDLINE, MeSH headings and free-text terms were combined with Boolean operators. The search string was: ("Chatbot" OR "Conversational Agent" OR "Virtual Therapist" OR "Digital Therapeutic") AND ("Cognitive Behavioural Therapy" OR "CBT" OR "Cognitive Behavioral Therapy") AND ("Efficacy" OR "Effectiveness" OR "User Engagement" OR "Adherence" OR "Dropout") AND ("Mental Health" OR "Depression" OR "Anxiety" OR "Well-being")

Reference lists of included systematic reviews^{1, 13, 24} and forward citation tracking of key RCTs^{2, 3} were used to identify additional eligible studies. Grey literature, commercial white papers, and app-store marketing materials were excluded because they rarely report attrition or sampling detail.

Studies from any country were considered, but only English full texts were included, which may miss trials published only in other languages despite local-language bots¹⁶.

2.3 Inclusion and Exclusion Criteria

Eligibility was assessed by study type, intervention focus, reported outcomes, language, and publication date.

We accepted RCTs and other peer-reviewed empirical reports, including longitudinal and mixed-methods studies, together with systematic or scoping reviews from which outcome numbers could be taken. We did not use editorials or opinion articles unless they reported original data, and we skipped conference abstracts that lacked full results, as well as papers that discussed theory without empirical findings.

The intervention had to involve a chatbot delivering CBT or CBT-based skills (e.g., cognitive restructuring, behavioural activation). Static workbook apps, clinician-led telehealth, and bots used only for triage or scheduling were excluded.

Papers had to report symptom outcomes, engagement or attrition, and/or access-related findings. Studies focused only on NLP model performance without clinical endpoints were excluded.

Only English full-text articles were included. Publications before January 2018 were excluded.

2.4 Study Selection Process

Records were deduplicated, then screened by title and abstract. Potentially eligible full texts were read in full. Articles were dropped if the tool was not a CBT chatbot, if relevant outcomes were missing, or if chatbot effects could not be separated from intensive human support.

Two reviewers independently screened titles and abstracts; disagreements were resolved by discussion. Full-text review followed the same process. Once full texts had been reviewed, 24 studies met all criteria and were carried forward into the synthesis.

A simple log recorded exclusion reasons at full-text stage (wrong intervention, missing outcomes, or inability to separate chatbot from intensive human support). This supported transparency but was not published as a PRISMA flow diagram because the review was structured rather than systematic.

2.5 Data Extraction

Extracted fields were chosen to fit the three aims of the review. For symptoms, we noted who was studied, which instruments were applied (PHQ-9, GAD-7, and similar scales), how long follow-up lasted, what the comparison arm was, and whether the authors reported effect estimates. Engagement items included retention, dropout, session counts, and alliance scores when reported (e.g., WAI-SR). Access-related notes covered participant background, income or education where stated, comfort with technology, and authors' comments on barriers to use.

Where studies reported both intention-to-treat and per-protocol results, we prioritised intention-to-treat estimates. Platform name, country, and whether the bot was rule-based or generative were recorded to support comparison across products²¹.

2.6 Thematic Synthesis

Results were grouped into three themes: clinical efficacy of chatbot-delivered CBT; user engagement, adherence, and dropout; and equity of access and structural barriers to use.

Within each theme, studies were compared by population (e.g., students, chronic disease cohorts, health workers), intervention duration, and architecture type. Contradictory findings were retained rather than resolved by vote counting. The synthesis emphasises patterns reported in more than one study and notes where evidence rests on a single trial.

Quality appraisal was narrative rather than scored. We noted obvious risks such as open-label delivery, short follow-up, selective reporting of engagement, and industry involvement in trial design^{5,21}. These features are common in digital mental health and do not automatically invalidate findings, but they caution against treating any single RCT as definitive proof of effectiveness for routine commissioning.

We did not score formal risk-of-bias tools study by study. Many trials were open-label, short, and inconsistently reported engagement^{5,6}; we note this when weighing the findings below.

No statistical meta-analysis software was used to recompute pooled effects. Where primary studies or prior reviews reported pooled estimates, we cite those values without repeating the pooling calculations.

When authors reported multiple follow-up points, we recorded the longest available post-intervention estimate consistent with the study's primary analysis plan. If a trial reported both completer-only and intention-to-treat analyses, we retained the intention-to-treat figure for the synthesis narrative⁵. Studies that bundled chatbot use with other digital modules were included only when chatbot-specific effects or clearly separated engagement data could be identified; otherwise they were excluded at full-text stage. This conservative rule may omit some hybrid digital programmes but reduces the risk of attributing blended interventions solely to conversational agents.

3. Results

3.1 Structural Overview of the Evidence

The reviewed studies describe two main types of CBT chatbots: rule-based systems and newer generative architectures¹². Rule-based platforms make up most commercially validated tools, including Woebot, Wysa, and Youper. These systems typically use structured decision trees developed with clinical input. They also limit open-ended input so that conversations stay within predefined therapeutic pathways². Newer generative systems, such as Limbic Care, use large language models to support more natural dialogue and to adjust CBT tasks during conversation³.

Samples ranged from stressed university students¹⁵ to people with arthritis or diabetes¹⁷, so chatbot CBT has been tested both for subclinical users and for chronic disease cohorts. Evaluations during COVID-19 among health workers¹⁸ and in rural India¹⁰ show that evidence is not limited to wealthy student groups.

Table 1 lists five commonly cited platforms and the main published findings linked to each. Other bots appear in included reviews^{21, 24} but with less trial data.

Table 1. Overview of selected CBT chatbot platforms, architecture types, primary clinical targets, and principal evidence cited in the reviewed literature.

Chatbot Platform	Underlying Architecture	Primary Clinical Targets	Evidence Base Highlight
Woebot	Rule-based / NLP	Depression, Anxiety, Postpartum	Reported symptom gains in RCTs; lower dropout than static iCBT in some comparisons ^{2,13} .
Wysa	Mixed (rule-based scripts + AI detection)	Anxiety, Depression, Chronic Pain	Reported benefit in chronic illness populations; relatively strong engagement ¹⁷ .
Youper	Rule-based / AI personalization	Depression, Anxiety	Longitudinal Youper data linked symptom change to repeated successful emotion-regulation attempts ¹¹ .
Limbic Care	Generative AI (LLM)	Depression, Anxiety	2.4x higher usage frequency and 3.8x longer use duration than static CBT workbooks ³ .
Fido	Rule-based (Polish language)	Subclinical Depression, Anxiety	Reported symptom improvement in a Polish subclinical sample; evidence for non-English delivery ¹⁶ .

3.2 Efficacy of Chatbot-Delivered CBT: Symptom Outcomes and Effect Sizes

Published studies suggest that chatbot-delivered CBT is associated with reductions in depression and anxiety symptoms, especially over short intervention periods of two to eight weeks¹. In trials with waitlist, information-only, or static workbook controls, chatbot arms often showed similar or better symptom change.

Outcome measures varied across trials. PHQ-9 and GAD-7 were common in student and general adult samples^{15, 1}, while condition-specific scales appeared in chronic disease cohorts¹⁷. Despite this variation, most reports described symptom improvement in chatbot arms over roughly two to eight weeks.

A 2025 meta-analysis in adolescents and young adults reported benefit on distress and related behaviours¹. Adult short-course meta-analyses report small-to-moderate gains on depression and anxiety scales¹³, though effect sizes shift with severity, comparator, and follow-up length.

Observational data from commercial platforms support these trial findings. In a longitudinal Youper cohort (N=4,517), anxiety and depression scores fell in the first two weeks of use, with Cohen's d values of

0.57 and 0.46 respectively¹¹. Authors linked more frequent successful emotion-regulation attempts during use with wider symptom improvement, describing a cumulative regulation pattern¹¹. Woebot trials also reported early reductions in depressive symptoms, together with greater user involvement in day-to-day self-management².

These tools may therefore be relevant not only in general subclinical samples but also among people managing physical comorbidity. In an RCT among adults with arthritis or diabetes, Wysa users showed significant reductions in depression and anxiety severity over four weeks compared with a no-treatment control¹⁷. This suggests that chatbot CBT may also help with psychological distress linked to chronic physical illness²².

Chronic disease samples raise a distinct clinical question: whether chatbots mainly reduce general distress or also support self-management tasks such as medication adherence and pain coping. Available Wysa-related reports emphasise mood symptoms more than disease-specific behavioural endpoints¹⁷, so the precise mechanism in comorbid populations remains partly unclear.

The available evidence is also not restricted to English-language apps. In a trial of Fido, a Polish-language CBT chatbot, subclinical young adults showed reductions in anxiety and depression similar to those reported for major English-language platforms¹⁶. PracticePal, tested in rural India, reported feasibility and early symptom change when a conversational agent supported brief behavioural activation¹⁰. Fido in Polish¹⁶ and PracticePal in rural India¹⁰ show that local language and delivery context matter for feasibility, not only translation of English content.

University-based RCTs add further detail. A Chinese student trial reported moderation by financial stress, implying that chatbot benefit may be uneven across socioeconomic subgroups even within a single campus sample¹⁵. Such heterogeneity cautions against treating student trials as universally generalisable.

Reviews of CBT chatbots stress reading effect sizes next to adherence and user satisfaction¹². Where several trials exist for one platform, short-term symptom gains look similar, but different comparators make direct ranking unreliable^{13, 21}.

Comparator choice strongly shapes how large effects appear. Waitlist controls often inflate apparent benefit because inactive comparison groups receive no attention placebo^{1, 13}. Trials against active digital workbooks or brief psychoeducation show smaller margins, which is more relevant for commissioners deciding between already-available low-intensity options^{3, 6}. Head-to-head studies between named chatbot products remain rare, so services cannot yet rank platforms on independent efficacy grounds alone.

3.3 User Engagement, Adherence, and Dropout Patterns

Although clinical benefits have been reported, keeping users engaged remains a major challenge in digital mental health. Interactive chatbots were introduced partly because static iCBT programmes often had poor adherence. Conversation-based tools appear to sustain use better than passive apps, although dropout remains an important limitation.

One meta-analysis of chatbot RCTs reported intervention-group attrition of 21.84%⁵. Attrition was higher in stand-alone apps without human support, apps without a visible agent representation, and apps without integrated symptom tracking⁵. Because apps are simple to install, they are also easy to leave; users commonly disengage when early interactions do not feel useful or supportive⁶.

Studies defined dropout differently: some counted anyone who stopped opening the app, others used missed symptom questionnaires. The pooled 21.84% attrition figure should therefore be treated as a rough benchmark rather than a precise planning statistic for local services⁵.

Trials with optional human coaching or clinician signposting were not pooled with fully self-guided apps. Where human contact was available, retention often improved, although the extra labour cost reduces scalability^{6, 8}.

Some design features were associated with better retention. In the Limbic Care RCT, generative AI delivery was linked to higher engagement than static digital workbooks³. In an RCT of Limbic Care, engagement frequency was about 2.4 times higher and total use time about 3.8 times longer than in a digital CBT workbook comparator³. Generative systems may also tailor homework more closely to the user's current barriers, which can make CBT tasks feel more relevant than generic workbook content⁴.

Reminder notifications and streak-based rewards were common in commercial apps but were not consistently linked to better clinical outcomes in the reviewed trials⁶. Their main effect may be to increase opens during the first week rather than to sustain cognitive work in later weeks.

Engagement was also linked to the digital therapeutic alliance (DTA). Studies using the Working Alliance Inventory-Short Revised (WAI-SR) reported meaningful alliance scores with fully automated chatbots in app-based samples¹⁴. Qualitative diary work suggests that sustained use is related to fit between the user's communication style and the chatbot's tone, as well as whether responses are experienced as empathic and non-judgmental¹⁹. When users reported a stronger alliance, adherence appeared more stable, which supports the importance of conversational design in chatbot CBT.

A Wysa service evaluation with health workers during COVID-19 found use shaped by shift patterns and acute work stress¹⁸, settings that may suit short on-demand chats more than long homework blocks.

Dropout timing also varied. Several trials reported steepest attrition in the first seven to ten days, before users completed enough CBT modules to practise core skills^{5,6}. Onboarding flows that front-load empathic dialogue without clarifying homework expectations may therefore inflate early opens while doing little to sustain cognitive work. Where studies reported session-level data, users who continued beyond initial mood checks often had better symptom trajectories than early dropouts^{2,3}.

3.4 Equity of Access: Digital Disparities and Structural Barriers

Chatbots are often promoted as highly accessible because they can be used outside clinic hours and may lower stigma around help-seeking⁹. In practice, access is still uneven: connection quality, devices, literacy, and regulation all influence who can use these tools.

A basic digital divide remains central. Use often depends on reliable internet, a smartphone, and basic digital skills—conditions more commonly met by younger, urban, and socioeconomically advantaged users²³. Groups who might benefit most from low-barrier support, including older adults, rural residents, and people with fewer resources, are often less able to use these tools consistently.

Algorithm design may create additional inequity. Models trained mainly on Western, Educated, Industrialized, Rich, and Democratic (WEIRD) samples may not transfer equally well to other populations²³. Historical diagnostic disparities can also be reflected in automated triage or symptom detection, which may reduce the relevance of chatbot responses for some users²³. If culturally specific expressions of distress are missed, the intervention may feel inappropriate or ineffective.

Regulation may widen these gaps further. As agencies such as the FDA require stronger clinical validation and assign designations such as Breakthrough Device to selected platforms, some evidence-based tools move toward prescription-based use within formal healthcare systems⁷. People already connected to care may therefore gain access to validated products, while uninsured or marginalized users may rely on unregulated commercial apps⁷. As a result, people with the least access to formal care may be left with weaker or unregulated app options rather than the best-validated tools.

Intersectional work on digital mental health shows that gender, ethnicity, income, and disability often overlap²³. Chatbots can still favour users who already have devices, data plans, and confidence with apps unless services fund outreach and support.

Older adults appear infrequently in chatbot RCTs²². When they participate, complex interfaces and weak onboarding often limit use. Larger text, simpler menus, or voice input could help but are rarely tested in published trials.

Low- and middle-income settings face additional constraints on mobile data cost and device sharing within households^{9,10}. A chatbot that works in a funded university trial may still be impractical for users who rely on prepaid data or shared phones.

3.5 Cross-Cutting Patterns

Several patterns cut across the three themes above. Rule-based bots account for most trials, while generative systems show stronger early engagement signals^{3,21}. Symptom gains are better documented than long-term functioning or recovery. Marketing claims about access often run ahead of equity data in peer-reviewed work^{7,23}.

Few included studies reported return-to-work, sustained remission beyond twelve weeks, or other functional endpoints. Short-term mood improvement is therefore the most defensible general conclusion from the current evidence base.

Implementation context varied between studies in ways that are easy to overlook when reading headline effect sizes. Some trials recruited through university counselling portals¹⁵, others through employer wellness programmes or primary-care waiting lists^{8,17}. Where recruitment went through a clinician, baseline severity and motivation may differ from self-initiated app-store downloads. Few studies compared chatbot outcomes

across these entry routes within the same product, so it is difficult to say whether symptom gains would replicate if the same bot were offered purely as a public download.

Safety reporting was inconsistent across included trials. Several commercial platforms describe automated crisis detection and signposting to helplines in product documentation, but peer-reviewed papers did not always report adverse events, crisis escalations, or inappropriate advice detected during study periods^{20,12}. Where large language models allow open-ended dialogue, the risk of harmful or misleading responses may depend on guardrail design as much as on the underlying CBT curriculum. This gap limits confident comparison of rule-based and generative systems on safety grounds, even when engagement metrics favour generative designs.

4. Discussion

4.1 Chatbot CBT as a Socio-Technical Intervention

These findings support viewing CBT chatbots as socio-technical interventions rather than simple software delivery systems. In this setting, outcomes are shaped not only by CBT content but also by interface design, wording, and whether the app fits into everyday use¹⁹.

The idea of a digital therapeutic alliance captures this point. In face-to-face CBT, progress depends on cooperation between therapist and patient; with a chatbot, the software itself becomes part of that working relationship¹⁹. Timing of responses, reflective prompts, and conversational tone may matter clinically as much as the CBT technique being delivered. Platforms using generative AI can also make interactions feel more responsive, which may increase session frequency and overall use³. For that reason, evaluation of chatbot CBT should include human-computer interaction measures alongside symptom scales.

Symptom scales alone cannot show whether users practised skills between chats. Brief homework completion checks or ecological momentary assessment may therefore be as informative as PHQ-9 or GAD-7 change in later implementation studies^{4,6}.

Where a bot is placed also matters. Self-help after GP referral, employer wellness schemes, and campus counselling each create different uptake constraints^{18,15}. Trials alone cannot show how a tool will perform once reimbursement and local support are fixed.

4.2 The Engagement Paradox: Accessibility vs. Sustained Use

The results point to a familiar tension in digital mental health. Features that make chatbots easy to access—availability on personal devices, low entry barriers, anonymity, and self-paced use—may also make it easier for users to stop⁵. In face-to-face therapy, cost, travel, scheduling, and accountability to a clinician create friction that can support attendance; leaving a therapy room is socially and practically harder than closing an app.

Developers therefore need to balance features that encourage opening the app with support for the more demanding parts of CBT work. Notifications, mood-tracking streaks, and highly agreeable dialogue may increase short-term opens, but they do not automatically produce the cognitive and behavioural change that CBT requires⁴. Longer-term value may depend less on surface engagement metrics such as daily active users and more on whether users complete homework, practise reframing, and apply skills outside the app³. Homework completion was rarely the primary endpoint in included trials, even though CBT theory treats between-session practice as central to change.

Generative systems may help here by adapting tasks when users struggle, rather than allowing disengagement when therapy becomes uncomfortable⁴.

Clinicians and commissioners should read engagement figures with caution. High login counts may mean curiosity or distress, not therapeutic work. Session logs are more informative when paired with symptom change and homework completion⁶.

Hybrid models are worth testing. Chatbots that triage into human care can widen entry without replacing clinicians⁸, which may suit systems with a duty-of-care obligation.

Cost and workforce implications deserve equal attention. Chatbots are often framed as saving clinician time, yet hybrid models that combine automated dialogue with human check-ins reintroduce labour costs that must be weighed against any reduction in waiting lists^{6,8}. If uptake concentrates among users who would otherwise have self-managed without formal care, net system benefit may be smaller than headline accessibility claims suggest.

4.3 Who Gets Left Behind: Intersectionality of Digital Exclusion

Available evidence suggests that users who already have stronger digital and socioeconomic resources may benefit most from chatbot-delivered CBT. Exclusion is usually the result of several overlapping constraints, including connectivity, digital skill, income, and how care is organized.

Language is another access issue. Many early chatbots were built first for English-speaking markets, even though some of the largest unmet need is reported in low- and middle-income countries⁹. The Polish-language Fido trial shows local-language bots are feasible¹⁶, but examples remain few. NLP performance in highly inflected languages or regional dialects is seldom reported in English-language efficacy papers.

Gender, age, disability, and rural residence can overlap²³. A middle-aged woman with chronic illness and limited smartphone experience may download an app yet still be unable to use it comfortably^{17, 22}. Outreach, digital skills help, and clinician signposting may be needed rather than assuming users will self-enrol.

Rural users may face weak mobile coverage as well as long travel to any fallback clinic²³. Chatbots do not remove those barriers unless low-bandwidth or offline modes are designed in.

4.4 Regulatory and Governance Implications

The shift from rule-based scripts to generative large language models has raised new governance questions. A further concern with generative models is hallucination, where the system may produce fluent but incorrect or unsafe therapeutic content¹². In crisis situations, such errors could carry serious clinical risk²⁰.

Regulators still lack a settled framework for classifying and overseeing these products. The funneling problem described earlier is relevant here: treating all chatbots as prescription-only medical devices may improve oversight for some products while reducing access for others⁷. A tiered approach may be more appropriate. Lower-intensity wellness or psychoeducational bots could remain broadly available with basic safety requirements, such as crisis detection and escalation pathways. More advanced generative tools intended for psychiatric triage or treatment support may need stronger pre- and post-market review, including monitoring for drift and bias¹². Data privacy also remains central, especially where sensitive emotional content could be commercialized without clear user consent.

Policy commentaries call for regulation that protects users without blocking useful tools⁷. At minimum, products sold as CBT support should state training data limits, clinical boundaries, and crisis escalation steps.

The same app may count as a wellness product in one country and a regulated device in another^{7, 20}. Users can then face different safety standards depending on where they download it, which makes clear labelling and crisis routing important.

Clinics reviewing procurement should check whether marketing claims about CE marking, FDA clearance, or equivalent status match the population they serve^{7, 20}. Evidence from US student trials does not automatically transfer to older primary-care patients in European systems.

4.5 Limitations of This Review

This structured narrative synthesis has several limitations. Peer-reviewed literature often lags behind commercial product updates, especially for large language model-based chatbots. Most included studies come from Western or East Asian settings and evaluate short interventions, typically two to eight weeks. Finally, this review does not perform formal pooled meta-analysis; effect estimates should therefore be interpreted at the level of individual studies.

Positive trials may be overrepresented in the literature. Attrition and equity are reported less often than symptom scales, which limits comparison across studies^{5, 6}. The English-only inclusion criterion may also miss relevant non-English evaluations.

Finally, rapid product updates—especially for large language model backends—mean that published trial evidence can become outdated within months. Readers should treat platform descriptions as snapshots of the literature rather than live product specifications.

We also did not contact developers for unpublished subgroup analyses or updated product versions. Commercial chatbots change frequently, and trials often evaluate earlier builds than those currently distributed²¹. Readers comparing platforms should therefore check whether cited evidence matches the version patients would download at the time of implementation.

4.6 Practical Recommendations

Health services may use chatbot CBT as a low-intensity stepped-care step if escalation rules for risk and complexity are clear^{8,20}. Commissioners should ask for attrition figures and independent safety review before wide rollout.

Developers should invest in alliance-oriented design and culturally adapted content, not only model upgrades^{14,19}. Generative dialogue needs guardrails and audit logs.

Head-to-head trials of rule-based and generative bots, longer follow-up, and uptake broken down by socioeconomic group are still scarce²³.

Managers should check whether chatbot contracts absorb budget without shortening waitlists. One practical indicator is whether signposting reduces time to first clinician contact when escalation is needed⁸.

Schools and employers introducing chatbot mental health tools should set clear boundaries about data use, duty of care, and when a student or employee is referred onward^{18,15}. Without those rules, voluntary apps may collect sensitive content without a linked support pathway.

Commissioners weighing chatbot procurement should treat implementation as a service redesign question rather than a single product purchase. Integration with existing referral pathways, data-sharing agreements, and staff training on escalation all influence whether users who deteriorate or disclose risk receive timely human follow-up^{8,20}. A platform that performs well in a standalone RCT may still fail in routine practice if primary-care staff are unsure when to recommend it or how to interpret usage alerts. Service-level pilots with pre-specified safety monitoring may therefore be more informative than additional short efficacy trials on similar populations.

4.7 Future Research Directions

Hybrid models that add brief human check-ins to chatbot CBT deserve further study because they may cut dropout while keeping some scalability⁶. We also need implementation work in NHS-style systems, employer programmes, and schools—not only stand-alone RCTs.

Shared reporting of engagement dose, safety incidents, and who actually uses each product would make cross-study comparison easier²¹. Independent academic follow-up of commercial updates will matter as regulatory categories shift.

Trials that keep human support constant while swapping rule-based and generative architectures could separate conversational design from CBT content effects^{12,3}, helping clinicians decide what to recommend in routine signposting.

Implementation research should track how chatbots are offered in routine care: opt-in self-help, post-GP referral, or embedded in chronic disease management^{22,17}. Uptake and outcomes may differ sharply across these pathways even when the same app is used.

Equity research on chatbot CBT remains thin relative to the volume of symptom-outcome trials. Few studies stratified outcomes by education, income, or prior digital mental health use, even though these factors plausibly shape both uptake and benefit^{23,15}. Future evaluations should report who enrolled, who completed, and who was excluded by design features such as smartphone-only access or English-only interfaces. Without that breakdown, claims that chatbots democratise access risk overstating reach among groups already underserved by conventional care.

5. Conclusions

Delivering CBT through chatbots has moved from early feasibility studies toward products used alongside, or instead of, clinic-based care. Across the studies summarised in this review, scripted and generative conversational agents were linked to symptom reductions in anxiety and depression among subclinical users and people living with chronic conditions. By supporting CBT exercises through dialogue, they may reduce some geographic and stigma-related barriers associated with clinic-based care.

Even so, population-level impact will turn on retention and on who can actually use these services. Generative designs may improve engagement by personalising between-session tasks, but technology alone cannot remove structural inequality. Measures to improve digital literacy, embed inclusive design, and regulate products fairly may help ensure that AI mental health tools do not simply mirror existing inequities.

The field still needs longer follow-up, cross-cultural validation, and governance models that keep effective chatbots within reach of underserved groups. Used selectively within stepped-care pathways and

monitored for safety and equity, CBT chatbots may offer a scalable complement to human-delivered therapy; they are unlikely to replace the need for trained clinicians in complex or high-risk cases.

Future monitoring should combine symptom trends, service use, and equity indicators so that digital expansion can be judged against public health goals rather than download counts alone.

For researchers, the priority is no longer to show that chatbots can reduce questionnaire scores in controlled settings. The harder questions concern durability of benefit, safe escalation when risk emerges, and whether these tools widen or narrow gaps in who receives effective psychological support^{23, 8}. Answering them will require closer collaboration between clinical trialists, implementation scientists, and the services that decide whether chatbot CBT belongs in routine care pathways.

Clinical and policy leads should judge chatbot CBT by how it is implemented locally, not as a generic technical fix. That framing fits the mixed evidence on outcomes, retention, and access reported here.

Taken together, the literature positions chatbot-delivered CBT as a plausible low-intensity option with consistent short-term symptom signals, uneven engagement, and unresolved questions about equity and governance. The next phase of evidence will need to shift from proof-of-concept trials toward implementation studies that link conversational design, safety monitoring, and real-world access in the health systems that might adopt these tools.

Author Contributions:

Conceptualization: Author 1 (Alicja Kwiatkowska), Author 2 (Jakub Juskowiak)

Methodology: Author 3 (Mateusz Lis), Author 4 (Zuzanna Julia Komin)

Software: Author 5 (Maryan Liasota)

Validation: Author 6 (Eliza Zofia Kabat), Author 7 (Michalina Justyna Kozłowska)

Formal Analysis: Author 8 (Krzysztof Konopka), Author 9 (Maciej Kamil Majchrzak)

Investigation: Author 1 (Alicja Kwiatkowska), Author 10 (Faustyna Magdalena Kołodziej)

Resources: Author 2 (Jakub Juskowiak), Author 3 (Mateusz Lis)

Data Curation: Author 4 (Zuzanna Julia Komin), Author 5 (Maryan Liasota)

Writing – Original Draft: Author 6 (Eliza Zofia Kabat), Author 7 (Michalina Justyna Kozłowska)

Writing – Review & Editing: Author 8 (Krzysztof Konopka), Author 9 (Maciej Kamil Majchrzak), Author 10 (Faustyna Magdalena Kołodziej)

Visualization: Author 1 (Alicja Kwiatkowska), Author 2 (Jakub Juskowiak)

Supervision: Author 3 (Mateusz Lis), Author 4 (Zuzanna Julia Komin)

Project Administration: Author 5 (Maryan Liasota), Author 6 (Eliza Zofia Kabat)

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement: The authors declare no conflicts of interest.

Declaration of the use of generative AI and AI-assisted technologies in the writing process: In preparing this work, the authors used ChatGPT for the purpose of improving language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1. Feng, X., Tian, L., Ho, G. W. K., Yorke, J., & Hui, V. (2025). The effectiveness of AI chatbots in alleviating mental distress and promoting health behaviors among adolescents and young adults: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 27, e79850. <https://doi.org/10.2196/79850>
2. Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2), e19. <https://doi.org/10.2196/mental.7785>
3. McFadyen, J., Habicht, J., Dina, L.-M., Harper, R., Hauser, T. U., & Rollwage, M. (2026). Increasing engagement with cognitive-behavioral therapy (CBT) using generative AI: A randomized controlled trial (RCT). *Communications Medicine*, 6, 132. <https://doi.org/10.1038/s43856-025-01321-8>
4. Oewel, B., Areán, P. A., & Agapie, E. (2024). Approaches to tailoring between-session mental health therapy activities. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 1–19. <https://doi.org/10.1145/3613904.3642856>
5. Jabir, A. I., Lin, X., Martinengo, L., Sharp, G., Theng, Y.-L., & Tudor Car, L. (2024). Attrition in conversational agent–delivered mental health interventions: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 26, e48168. <https://doi.org/10.2196/48168>
6. Smith, K. A., Davenport, T., Ospina-Pinillos, L., Guille, C., Arroll, B., Patel, V., & Torous, J. (2025). Engagement and attrition in digital mental health: Current challenges and potential solutions. *npj Digital Medicine*, 8(1), 398. <https://doi.org/10.1038/s41746-025-01778-w>
7. Palmer, A., & Schwan, D. (2025). Digital mental health tools and AI therapy chatbots: A balanced approach to regulation. *Hastings Center Report*, 55(3), 15–29. <https://doi.org/10.1002/hast.4979>
8. Habicht, J., Viswanathan, S., Carrington, B., Hauser, T. U., Harper, R., & Rollwage, M. (2024). Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. *Nature Medicine*, 30(2), 595–602. <https://doi.org/10.1038/s41591-023-02766-x>
9. van der Schyff, E. L., Ridout, B., Amon, K. L., Forsyth, R., & Campbell, A. J. (2023). 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*, 25, e46448. <https://doi.org/10.2196/46448>
10. Agrawal, A., Weiss, H. A., Gupta, R., Pednekar, S., Wilkinson, T., Kirk, K., & Patel, V. (2025). A conversational agent (PracticePal) to support the delivery of a brief behavioral activation treatment for depression in rural India: Development and pilot-testing study. *JMIR Formative Research*, 9, e73563. <https://doi.org/10.2196/73563>
11. Birk, M. N., Darcy, A. M., & Chen, E. (2021). Acceptability and effectiveness of artificial intelligence therapy for anxiety and depression (Youper): Longitudinal observational study. *Journal of Medical Internet Research*, 23(6), e26771. <https://doi.org/10.2196/26771>
12. Im, C.-H., & Woo, M. (2025). Clinical efficacy, therapeutic mechanisms, and implementation features of cognitive behavioral therapy–based chatbots for depression and anxiety: Narrative review. *JMIR Mental Health*, 12, e78340. <https://doi.org/10.2196/78340>
13. Zhong, W., Luo, J., & Zhang, H. (2024). The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: A systematic review and meta-analysis. *Journal of Affective Disorders*, 356, 371–380. <https://doi.org/10.1016/j.jad.2024.04.057>
14. Beatty, C., Malik, T., Meheli, S., & Sinha, C. (2022). Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): A mixed-methods study. *Frontiers in Digital Health*, 4, 847991. <https://doi.org/10.3389/fdgth.2022.847991>
15. Wang, Y., Li, X., Zhang, Q., Yeung, D., & Wu, Y. (2025). Effect of a cognitive behavioral therapy–based AI chatbot on depression and loneliness in Chinese university students: Randomized controlled trial with financial stress moderation. *JMIR mHealth and uHealth*, 13, e63806. <https://doi.org/10.2196/63806>

16. Karkosz, S., Szymański, R., Sanna, K., & Michałowski, J. (2024). Effectiveness of a web-based and mobile therapy chatbot on anxiety and depressive symptoms in subclinical young adults: Randomized controlled trial. *JMIR Formative Research*, 8, e47960. <https://doi.org/10.2196/47960>
17. MacNeill, A. L., Doucet, S., & Luke, A. (2024). Effectiveness of a mental health chatbot for people with chronic diseases: Randomized controlled trial. *JMIR Formative Research*, 8, e50025. <https://doi.org/10.2196/50025>
18. Chang, C. L., Sinha, C., Roy, M., & Wong, J. C. M. (2024). AI-led mental health support (Wysa) for health care workers during COVID-19: Service evaluation. *JMIR Formative Research*, 8, e51858. <https://doi.org/10.2196/51858>
19. Xu, Z., Lee, Y.-C., Stasiak, K., Warren, J., & Lottridge, D. (2025). The digital therapeutic alliance with mental health chatbots: Diary study and thematic analysis. *JMIR Mental Health*, 12, e76642. <https://doi.org/10.2196/76642>
20. Yoon, S. C., An, J. H., Choi, J., Chang, J. H., Jang, Y. J., & Jeon, H. J. (2025). Digital psychiatry with chatbot: Recent advances and limitations. *Clinical Psychopharmacology and Neuroscience*, 23(4), 542–550. <https://doi.org/10.9758/cpn.25.1346>
21. Khosravi, M., & Izadi, R. (2026). Mental health chatbots and their technical features: A systematic review of reviews and a thematic analysis. *Global Mental Health*, 13, e10144. <https://doi.org/10.1017/gmh.2026.10144>
22. Hou, J., Lin, S., Teng, P., Han, Y., Luo, Y., & He, G. (2026). Application of chatbots in chronic disease management: A scoping review. *Digital Health*, 12, 20552076251411287. <https://doi.org/10.1177/20552076251411287>
23. Yesha, R., Orezza, M. C. E., Sims, K., & Landau, A. Y. (2026). Digital mental health through an intersectional lens: A narrative review. *Healthcare*, 14(2), 211. <https://doi.org/10.3390/healthcare14020211>
24. Bendig, E., Erb, B., Schulze-Thuesing, L., & Baumeister, H. (2022). The next generation: Chatbots in clinical psychology and psychotherapy to foster mental health—a scoping review. *Verhaltenstherapie*, 32(2), 64–76. <https://doi.org/10.1159/000501812>