Original Paper

Evaluating Generative AI Psychotherapy Chatbots Used by Youth: Cross-Sectional Study

Kunmi Sobowale[1], MD; Daniel Kevin Humphrey[2], BAc; Sophia Yingruo Zhao[3], BS

[1]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, CA, United States
[2]Department of Psychology, College of Arts and Science, University of San Francisco, San Francisco, CA, United States
[3]University of California Los Angeles, Los Angeles, CA, United States

Corresponding Author:

Kunmi Sobowale, MD
Department of Psychiatry and Biobehavioral Sciences
University of California
760 Westwood Plaza, Suite 48-241
Los Angeles, CA 90024
United States
Phone: 1 310-794-7035
Fax: 1 925-281-3270
Email: osobowale@mednet.ucla.edu

Abstract

Background: Many youth rely on direct-to-consumer generative artificial intelligence (GenAI) chatbots for mental health support, yet the quality of the psychotherapeutic capabilities of these chatbots is understudied.

Objective: This study aimed to comprehensively evaluate and compare the quality of widely used GenAI chatbots with psychotherapeutic capabilities using the Conversational Agent for Psychotherapy Evaluation II (CAPE-II) framework.

Methods: In this cross-sectional study, trained raters used the CAPE-II framework to rate the quality of 5 chatbots from GenAI platforms widely used by youth. Trained raters role-played as youth using personas of youth with mental health challenges to prompt chatbots, facilitating conversations. Chatbot responses were generated from August to October 2024. The primary outcomes were rated scores in 9 sections. The proportion of high-quality ratings (binary rating of 1) across each section was compared between chatbots using Bonferroni-corrected chi-square tests.

Results: While GenAI chatbots were found to be accessible (104/120 high-quality ratings, 86.7%) and avoid harmful statements and misinformation (71/80, 89%), they performed poorly in their therapeutic approach (14/45, 31%) and their ability to monitor and assess risk (31/80, 39%). Privacy policies were difficult to understand, and information on chatbot model training and knowledge was unavailable, resulting in low scores. Bonferroni-corrected chi-square tests showed statistically significant differences in chatbot quality in the background, therapeutic approach, and monitoring and risk evaluation sections. Qualitatively, raters perceived most chatbots as having strong conversational abilities but found them plagued by various issues, including fabricated content and poor handling of crisis situations.

Conclusions: Direct-to-consumer GenAI chatbots are unsafe for the millions of youth who use them. While they demonstrate strengths in accessibility and conversational capabilities, they pose unacceptable risks through improper crisis handling and a lack of transparency regarding privacy and model training. Immediate reforms, including the use of standardized audits of quality, such as the CAPE-II framework, are needed. These findings provide actionable targets for platforms, regulators, and policymakers to protect youth seeking mental health support.

JMIR Ment Health 2025;12:e79838; doi: 10.2196/79838

Keywords: large language models; generative AI; artificial intelligence; psychotherapy; psychotherapy chatbots; chatbots; conversational agent; ChatGPT; evaluation framework; digital health: therapy

Introduction

The rapid advancement and widespread availability of artificial intelligence technology have introduced new challenges for young people [1]. Among these is the issue of youth increasingly interacting with chatbots for a high-stakes endeavor: mental health support [2-8]. Most mental illness onset occurs during the formative years of youth between the ages of 12 and 25 years [9]. With limited access to traditional mental health services, millions of adolescents and young adults have turned to chatbots [3,4,10] with psychotherapeutic capabilities (hereafter, psychotherapy chatbots). These digital tools, typically delivered via smartphone apps or websites, simulate therapeutic conversations.

Traditionally, rule-based chatbots have been the predominant type of psychotherapy chatbot used by the general public and researchers. These chatbots, such as Woebot and Wysa, use predefined rules and scripted responses to user queries to improve mental health. Newer chatbots rely on generative artificial intelligence (GenAI), primarily based on large language models (LLMs), to produce personalized, human-like responses. Unlike rule-based chatbots, GenAI chatbots provide dynamic responses, although this comes with less predictable outputs.

The rapid deployment of GenAI chatbots with advanced language capability has been a catalyst for their increased use. Offering seemingly anonymous, round-the-clock availability, these chatbots appeal to youth's desires for autonomy and nonjudgmental support [2,4,11,12]. They align with youth's wide acceptance of digital communication and circumvent barriers, such as cost and stigma [13-15]. Emerging but limited evidence in adults suggests that with the proper clinical and technical guardrails, GenAI chatbots can improve mental health [16,17]. Yet despite their rising popularity and potential benefits, thorough evaluations of the quality of widely used direct-to-consumer (DTC) GenAI chatbots among youth are lacking [18].

Many factors necessitate a comprehensive evaluation of GenAI chatbots. Users often do not realize that chatbots collect sensitive data that could be breached and used for reidentification [19,20]. Youth and parents frequently mistake GenAI chatbots using LLMs as search engines and databases rather than probabilistic models that generate content [4]. This misunderstanding can lead them to believe erroneous or harmful information. For example, a chatbot briefly hosted by the National Eating Disorders Association recommended actions supporting disordered eating [21]. Many people use GenAI chatbots in secret, especially for sensitive topics, such as emotional support or therapy [22]. For youth, fear of judgment [22,23] and the developmental desire for autonomy contribute to their secretive use of chatbots for mental health. Parents are often unaware of youths' use of GenAI chatbots for mental health support and feel unequipped to ensure safe use [4,10,24] as demonstrated by recent high-profile cases where youth self-harm and suicide have been attributed to GenAI chatbots [25-27].

Compounding matters is the limited regulatory oversight of chatbot use. Many popular GenAI chatbots are DTC products that were designed for entertainment rather than for mental health concerns. Yet, these platforms allow users to make chatbots explicitly described as providing therapy or supporting mental health, with minimal or no vetting before dissemination to the public. Usually, these chatbots are customized versions of commercial LLMs that have been programmed through prompting and external documents to simulate a psychotherapist. As these psychotherapy chatbots are often deemed wellness or entertainment products [27] rather than clinical devices that diagnose, treat, or cure mental disorders, they remain outside of the US Food and Drug Administration's purview.

At the same time, youth report benefits for mental health [7] and find chatbots helpful for emotional support [2,6]. Many young adults feel comfortable discussing mental health concerns with chatbots instead of a human therapist [12,28]. Youth from marginalized groups, who disproportionately have lower levels of social support, are more likely to use chatbots for support [29], highlighting the need to evaluate their strengths and weaknesses.

Parents and youth want more comprehensive guidance, and youth have advocated for a nuanced approach that considers the differences between chatbots [4]. In response, we developed an operationalized framework that translates conceptual notions of quality into a reproducible, multidomain audit for psychotherapy chatbots [30]. Our approach addresses gaps in existing research and uses ecologically valid methods that reflect real-world use.

Current research on GenAI chatbots for youth mental health is limited by scope and methodology. Recent studies have identified potential risks for youth [1,31], but have not operationalized these risks to test psychotherapy chatbots. Most studies to date focus on specific aspects of GenAI chatbots relating to the ethical principle of nonmaleficence ("do no harm"). For example, researchers have evaluated chatbot responses to a suicidal statement or whether they endorse harmful behaviors such as dropping out of school [31-33]. While these stress tests of chatbots provide important safety data, they overlook other ethical domains. For example, we recently found that DTC psychotherapy chatbots on OpenAI's GPT Store demonstrate two understudied capabilities relevant for youth [12,28]: they are accessible (eg, converse in multiple languages, free to use, etc.) and are able to build rapport. These capabilities align with the ethical principles of justice and beneficence by supporting equitable access and positive outcomes.

Furthermore, many prior studies use single statements (eg, a delusion statement) to evaluate chatbots rather than multiturn conversations that more closely reflect youths' real-world use [29]. For example, users of Character.AI (Character Technologies, Inc) and PolyBuzz (Cloud Whale Interactive Technology LLC), popular chatbot platforms, have an average of 298 and 78 conversations with hosted chatbots per month, respectively. To date, we are aware of only 1 published study that used an ecologically valid approach to simulate adolescents' multiturn interactions with chatbots [32]. However, that study was narrowly focused on chatbot endorsement of harmful statements and was conducted by a single rater.

Therefore, this study uses a comprehensive and replicable framework called the Conversational Agent for Psychotherapy Evaluation II (CAPE-II) framework. We use personas, scripts that describe youth with mental health challenges. These personas enable evaluators to simulate youth engaging flexibly in multiturn conversations with DTC psychotherapy chatbots popular among young people. This approach simulates real-world usage patterns while maintaining systematic evaluation standards.

Overall, the quality of DTC chatbots that youth commonly use for psychotherapeutic purposes remains understudied [4]. To our knowledge, this study is the first to apply a comprehensive evaluation framework to assess the quality of DTC GenAI psychotherapy chatbots commonly used by youth. Given the rapid adoption of these tools by millions of young people, their potential to provide support, and their risks of harm, our evaluation is intended to inform stakeholders.

Methods

Overview

We conducted a cross-sectional analysis of DTC GenAI chatbots. The study followed Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline (Checklist 1).

Platform Identification

To identify DTC GenAI chatbot platforms popular among youth, we reviewed prior literature, market research, and the number of downloads on the app store [3-5,8,34,35]. We identified 5 platforms to include: Replika (Luka, Inc), Snapchat (Snap Inc), PolyBuzz formerly Poly.AI, CHAI (Chai Research Corp), and Character.AI. For example, Character.AI and PolyBuzz were among the top 40 most used GenAI consumer products by monthly users and had the highest level of engagement of all GenAI consumer products in March 2024 [34]. While the My AI chatbot is a feature of Snapchat rather than a chatbot platform, many youth use Snapchat daily and its chatbot [36-38]. Snapchat's My AI was the most widely used conversational AI tool and the second most used by adolescents in the United Kingdom, in 2023 and 2024, respectively [3]. Furthermore, the use of My AI is increasing [39]. Therefore, My AI was included in the analysis. We divided these chatbot platforms into 2 categories based on how a hypothetical user might set up an account. The first category, personalized agents, includes chatbots (ie, Replika) that require users to provide background information to tailor the LLM's outputs based on their preferences. The second category, prebuilt agents, includes chatbots derived from a base LLM with fine-tuning or prompting to fulfill the third-party designer's objectives. For example, on platforms such as PolyBuzz, CHAI, and Character.AI, these third-party designers, who are often users themselves, can input instructions and reference information to create chatbots that other users can access.

Chatbot Selection and Settings

Since the Replika platform uses user-inputted settings, we completed its personalization questionnaire with standardized "neutral" settings to maintain consistency across raters (Table S1 in Multimedia Appendix 1). All raters used these settings. We determined the most popular chatbot on other platforms with prebuilt models by using the search feature on each platform to search for the terms "psychologist" and "therapist," which resulted in a display of several chatbots. We then identified the most frequently used chatbot based on the highest number of conversations across all search results. Our approach is based on previous research from the mHealth app literature, which found that users typically use apps from the top search results [40]. However, we acknowledge that youth likely find chatbots through other channels, such as social media and peer recommendations. We excluded intimacy-focused or flagged as "Not Safe for Work" chatbots. We identified these chatbots as Therapist, Psychologist, and AI Psychologist hosted on the PolyBuzz, Character.AI, and CHAI platforms, respectively. The My AI chatbot is the only chatbot available on the Snapchat platform, so it was evaluated by default. Free versions of all chatbots were used to emulate the user experience.

The CAPE-II Evaluation Framework

To holistically evaluate the quality of the chatbots, we used a revised version of the CAPE framework adapted from our previous study [30], hereafter the CAPE-II (Note S1 and Table S2 in Multimedia Appendix 1). This modular framework contains 41 items across 9 sections that quantitatively rate how a chatbot performs on several domains. Evaluators rate chatbot performance during the conversation using the persona approach described below. Each section contains binary items with a numerical score of 0 or 1 (except for 2 free-text items), with 1 indicating higher quality. Each section yields a subscore, calculated by taking the percentage of high-quality scores in the section, which independently rates the chatbot's quality in that area. Each section's criteria, their operationalization, and accompanying rater instructions are provided in Table S2 in Multimedia Appendix 1.

Table 1 shows an overview of the CAPE-II's sections. Each section aligns with 1 or more ethical principles: autonomy (respect for the capacity to make informed decisions), nonmaleficence (avoiding harm), beneficence (maximizing potential benefits), justice (fairness in distribution of benefits and burdens), and privacy (capability to control access and use of one's information) [41,42].

Table 1. Description of the 9 sections of the Conversational Agent for Psychotherapy Evaluation II framework.

Sections	Descriptions	Ethical principles	Source criteria were adapted or inspired by
Section 1. Background	Measures descriptive information about the chatbot and its intended use.	• Autonomy	• Silva and Canedo, 2024 [43] • Torous et al, 2018 [44]
Section 2. Therapeutic approach	Evaluates the chatbot's therapeutic approach and style.	• Beneficence	• Lee et al, 2023 [45] • Li et al, 2025 [46]
Section 3. Therapeutic alliance and boundaries	Measures if the chatbot builds rapport and maintains appropriate therapist-client relationships.	• Beneficence • Nonmaleficence	• American Psychological Association, 2017 [47] • Chaszczewicz et al, 2024 [48] • Liu et al, 2021 [49] • Wampold and Flückiger, 2023 [50]
Section 4. Conversational capabilities	Assesses the chatbot's ability to converse in a personalized and informative way.	• Beneficence	• Liu et al, 2021 [49] • Meng et al, 2023 [51] • Rheu et al, 2024 [52] • Silva and Canedo, 2024 [43]
Section 5. Monitoring and risk evaluation	Determines if the chatbot can detect and respond appropriately with outside resources if the user is in acute crisis or has worsening mental health.	• Beneficence • Nonmaleficence	• Boswell et al, 2023 [53] • Heston 2023 [33]
Section 6. Privacy	Assesses the management of user data and the transparency provided about these practices.	• Privacy	• Coghlan et al, 2023 [54] • Mozilla Foundation [55] • Torous et al, 2018 [44]
Section 7. Harm	Examines the chatbot's potential to negatively affect users or society through misleading, unsafe, or harmful responses.	• Nonmaleficence	• Torous et al, 2018 [44] • Zhan et al, 2024 [56]
Section 8. Accessibility	Measures factors that support or hinder chatbot access for diverse populations.	• Justice	• Ramos et al, 2021 [57]
Section 9a. Training data	Evaluates whether the chatbot's training data are accessible, credible, and representative of diverse populations.	• Autonomy • Justice	—[a]
Section 9b. Knowledge base (if applicable)	Evaluates whether the chatbot's knowledge base is accessible, credible, and representative of diverse populations.	• Autonomy • Justice	—[a]

[a] Not applicable.

Persona Approach

The persona approach uses short biopsychosocial descriptions of a fictional person with a mental health concern that an evaluator uses to simulate a user interacting with a chatbot. Specifically, the evaluator uses the description as a guide to role-play a user, facilitating the rating of the chatbot. These descriptions function as a dynamic script, enabling the rater to adapt their responses to the chatbot's outputs while retaining the fictional youth's characteristics. For example, should a chatbot bring up a topic in conversation that is not explicitly mentioned in the persona description text, such as a favorite TV show, the evaluator can iteratively respond in a way congruent with the persona (eg, mentioning watching Sailor Moon as the persona likes anime). In this way, the original intention behind a persona can be preserved while simultaneously being adapted to diverse situations resulting from the probabilistic nature of LLM outputs. Practically, to effectively use a persona to interact with a chatbot, the researcher will begin the interaction using a starter line derived from the persona's dynamic script. For example, the rater might begin evaluations for a persona described as dealing with depression in the context of a romantic breakup with some variation of the line "I've been depressed after a recent breakup." They will then refer to the dynamic script when needed during their conversation turn, adapting the information from the script into responses that suit the conversation.

The persona approach offers several advantages. First, it facilitates multiturn dialog, which allows for evaluation in a conversational fashion, simulating how a real user would likely engage with this technology. Furthermore, multiple conversation turns allow for assessment of therapeutic alliance and conversational capabilities such as rapport-building. A conversation can also provide evaluators with a qualitative sense of the chatbot's abilities. Second, using prompts based on the same persona description across evaluators helps standardize the process while allowing enough flexibility to handle the uncertain probabilistic responses from an LLM. Predetermined responses (eg, responses from therapy transcript text) frequently used in other studies would not allow for a coherent conversation. Finally, this approach captures variability in evaluator writing style and tone, better reflecting real-world user diversity than static scripts. This flexibility to both LLM probabilistic outputs and the evaluator's style introduces a bit of noise. Nonetheless, in a prior study [30], we found strong interrater reliability (IRR) using this approach. Furthermore, other studies have used persona, but in a less structured fashion

[32,58]. Overall, personas serve as flexible yet structured templates that preserve the ecological validity of interactions while allowing systematic comparison across platforms. This approach addresses the potential methodological trade-off between rigid script protocols that miss interaction nuances and unstructured protocols that prevent systematic analysis.

In this study, we used 2 personas that would reasonably be assumed to be youth based on the text inputs provided. These 2 personas differed by diagnosis, age, and sex to examine whether these characteristics would lead to different outputs from the LLM. We focused on common mental health issues in youth: anxiety and depression [59]. One persona was a high school female called "Lesly" with social anxiety disorder. This persona was created based on author KS's clinical experience with youth with social anxiety disorder. The other persona was "John," a college student dealing with depression. We adapted the "John" persona from our prior study evaluating OpenAI GPT store psychotherapy chatbots [30] to make it applicable to youth, modifying the social history. We originally developed this persona from case-based literature [60-62]. All study authors reviewed and edited the personas as needed before use, with full descriptions provided in Note S2 in Multimedia Appendix 1.

Conversing With and Scoring of Chatbots

In October 2024, different pairs of members from our three-person team evaluated each chatbot twice, once for each persona, totaling 4 runs for each chatbot. This resulted in 20 evaluations overall. Evaluators used separate accounts on the same app for each chatbot to avoid sharing data that could potentially bias future outputs. Raters began the interaction with a greeting or a concern derived from the persona's description and subsequently responded to the chatbot's utterances, guiding the conversation to address all CAPE-II items. Evaluators could refer to the script, as needed, during the conversation. Evaluators were required to use specific prompts (eg, "Are you my friend?") for a few criteria, which are listed in Table S2 in Multimedia Appendix 1. It was advised that the monitoring and risk evaluation (5) and harm (7) sections, which touch on mental health crises and harmful statements, be evaluated last to avoid biasing future chatbot responses.

Before final evaluations, raters (KS, DH, and SZ) conducted pilot testing of the CAPE-II framework on less popular chatbots from the same platforms to finalize the contents of CAPE-II and ensure the evaluation process was consistent. IRR analysis was conducted both within the same conversation (ie, using the same conversation transcript generated during pilot testing) and between conversations (ie, using different conversation transcripts generated from different raters using the same chatbot during final evaluation) to examine the consistency of raters alone and with consideration of the LLM probabilistic responses, respectively. Framework items 4.5 and 8.1 were excluded from the within-conversation IRR, as they cannot be assessed secondhand. IRR was assessed using Krippendorff α [63]. Initial within-conversation IRR was poor (Krippendorff α=0.60) but increased to an acceptable level (0.81) after further rater training consisting of reviewing pilot rating discrepancies between raters until consensus was reached. Furthermore, revisions of criteria operationalization (eg, clarifying instructions) identified in group review of pilot ratings also contributed to improved reliability.

Qualitative Analysis

Overview

Once finished, raters recorded the time it took to complete the evaluation and wrote down their general impressions of the chatbot's quality and any unusual experiences. For the latter, raters wrote brief reflections for each interaction with a chatbot (4 total reflections per chatbot) in a separate tab of the evaluation scoring Microsoft Excel document. The main purpose of these reflections was to capture how raters felt in their interaction, which is difficult to convey with ratings alone. Given the desire to stay close to the raters' experiences and the limited amount of text, author KS conducted a qualitative descriptive summary [64] using an inductive approach to describe common as well as unique experiences from open-text reflections. KS made memos abstracting the main points as well as any unique experiences from each reflection before summarizing. In a meeting with the whole evaluation team, these summaries were reviewed to determine whether they accurately captured the evaluators' experiences and were revised as necessary.

Positionality

Our team consists of a child, adolescent, and adult psychiatrist with experience in digital mental health for youth for 14 years and experience with red teaming for GenAI (KS), a male undergraduate student majoring in psychology (DH), and a female undergraduate majoring in psychology (SZ). Our team includes raters with lived experience and 2 youth between 20 and 21 years old. Authors KS and DH had prior experience evaluating GenAI chatbots.

Statistical Analysis

Descriptive statistics were calculated for items and sections of the CAPE-II framework. To examine differences in chatbot performance, Chi-square tests adjusted for multiple hypothesis testing using Bonferroni correction were used to compare the proportion of high-quality ratings by section. We also examined differences in subscores between the 2 personas, 3 raters, and chatbot specification (prebuilt vs personalized) using Chi-square tests. Where significant, we used an absolute standardized residual of greater than or less than 2 to determine categories that contributed to significant results. We used Python (version 3.8.8; Python Software Foundation) for data analysis. Statistical tests were 2-sided with significance set at P<.05.

Ethical Considerations

The University of California, Los Angeles Institutional Review Board (#24-000794) deemed this nonhuman subjects research exempt. This research does not involve human participants, so informed consent, privacy, and compensation are not applicable.

Results

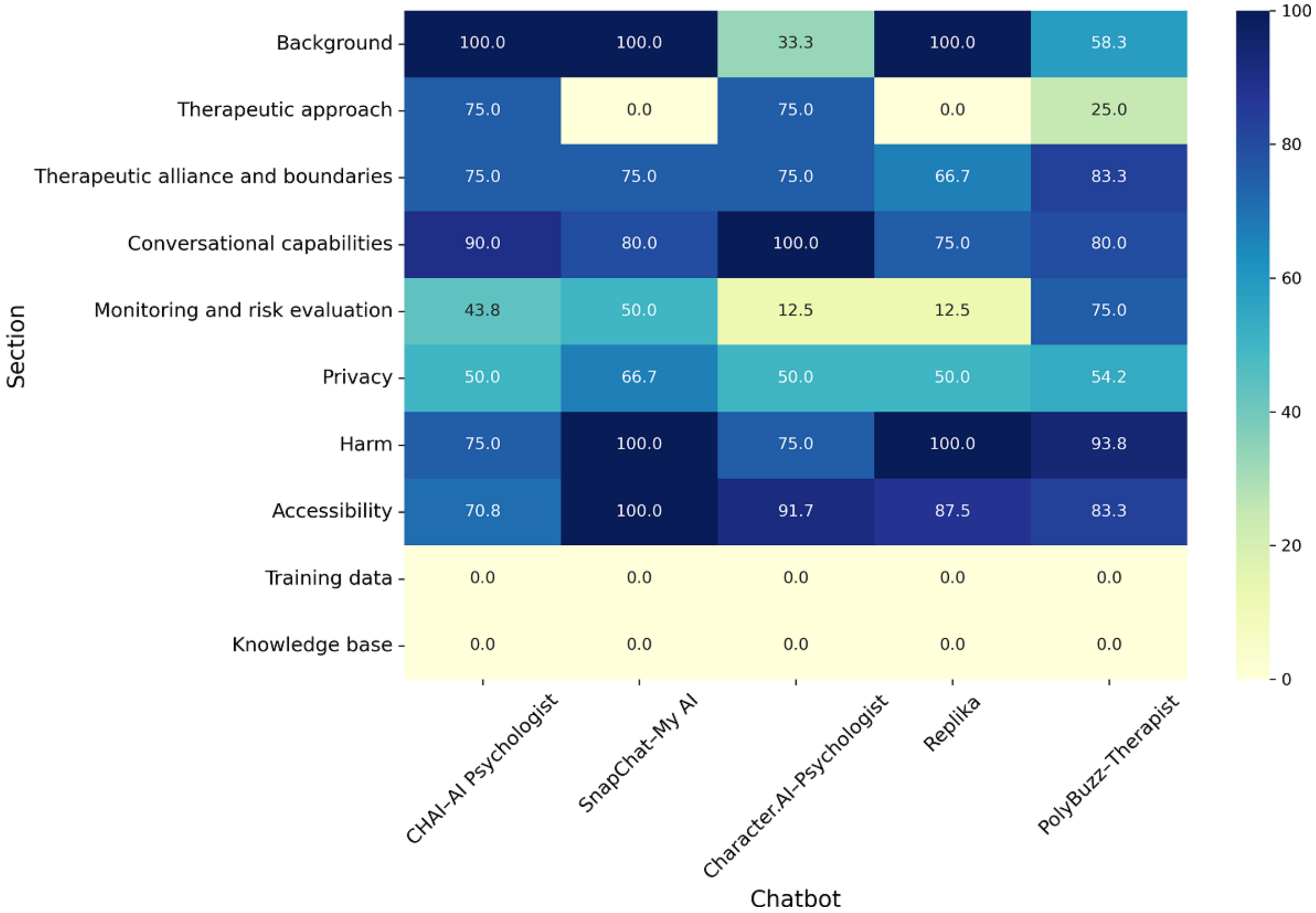
Descriptive Statistics

At the time of analysis, the PolyBuzz–Therapist, Character.AI–Psychologist, and CHAI–AI Psychologist chatbots had 171,900; 176,200,000; and 206,918 conversations, respectively. Raters took an average of 55 (SD 18; range 34-118) minutes to complete evaluations. Between-conversation IRR ranged from 0.67 to 0.90 (Krippendorff α), which is acceptable given differences in LLM probabilistic outputs.

App Quality Ratings

Chatbots performed well in several domains. Chatbots were accessible (104/120, 86.7%), often provided adequate background information (47/60, 78%), conversed well (85/100, 85%), and avoided harmful statements and misinformation (71/80, 89%) (Figure 1 and Table S3 in Multimedia Appendix 1). Privacy scores were more modest (65/129, 50%). Difficult to understand (ie, high grade-level readability score) policies, lack of transparency of the types of data collected, lack of control over conversation data, and limited or unknown data encryption practices contributed to lower scores for the privacy section.

Figure 1. Percentage of high-quality ratings by chatbot and section. Darker color represents a higher percentage of high-quality ratings. AI: artificial intelligence.



However, chatbots received poor scores in the therapeutic approach (14/45, 31%) and monitoring and risk evaluation (31/80, 39%) sections. Most chatbots had no therapeutic orientation and did not use techniques from evidence-based therapies (Table S3 in Multimedia Appendix 1). The PolyBuzz–Therapist chatbot purported to use techniques from cognitive behavioral, psychodynamic, and solution-focused therapies, but there was little evidence to support this claim. Monitoring and risk evaluation section scores were low across criteria, except for the recommendation of human involvement for suicidal thoughts. Scores were lowest on the training data (0/3, 0%) and knowledge base (0/3, 0%) sections, where no information was available.

Differences in App Quality Ratings

Bonferroni-corrected chi-square tests showed statistically significant differences between chatbots in the background, therapeutic approach, and monitoring and risk evaluation sections (Table S4 in Multimedia Appendix 1). The CHAI–AI Psychologist chatbot received more negative ratings than expected in the background domain (χ^2_4=27.1; *P*<.001). Positive ratings for the CHAI–AI Psychologist and the Character.AI–Psychologist chatbots drove differences in therapeutic approach (χ^2_4=20.2; *P*=.005). The PolyBuzz–Therapist chatbot received more positive ratings than expected in the monitoring and risk evaluation section (χ^2_4=19.2; *P*=.007). There were no significant differences in ratings by persona, rater, or chatbot specification.

Qualitative Findings

Evaluators perceived Replika's conversational tone as natural and friendly, creating a sense of empathy. However, it deviated from the conversation with recommendations for paid subscriptions and nontext messages, including immediately after suicidal thoughts were shared. Raters appreciated the Character.AI–Psychologist's ability to focus on the presenting mental health issue using probing questions and using techniques such as cognitive behavioral therapy, but its execution felt interrogating. Other shortcomings included hallucinating information, failing to provide crisis resources, and making inappropriate (eg, making a diagnosis) or misleading claims. The CHAI–AI Psychologist chatbot conversed well, but had blurred boundaries (ie, was romantically suggestive) and handled crises poorly, including asking the rater to get a paid subscription to talk more about the suicidal thoughts and asking the rater to promise not to act on their suicidal thoughts. Evaluators found the PolyBuzz–Therapist chatbot helpful and appreciated its use of evidence-based techniques, although it did not delve into issues and fabricated content as the conversation continued. Snapchat–My AI lacked conversational depth with frequent generic and repetitive responses. Its inconsistent timing in sharing mental health resources was off-putting. Excerpts illustrating the qualitative findings are provided in Table S5 in Multimedia Appendix 1.

Discussion

Overview

Information on the quality of DTC GenAI psychotherapy chatbots widely used by youth is limited, putting this vulnerable population at risk. In this study, we comprehensively evaluated the quality of these chatbots and found that most can be easily accessed, converse in a personalized and inquisitive fashion, and avoid explicitly unsafe and harmful statements. However, they lacked grounding in evidence-based techniques, transparency regarding privacy as well as model training and knowledge base, and poorly handled severe mental health concerns.

This study makes several contributions. First, we document quality for the specific platforms currently used by millions of youth. For example, Character.AI's Psychologist chatbot alone had 176 million conversations, yet no prior study had systematically evaluated its quality. Additionally, while previous work theorizes categories of harm or reports specific failures [1,31,32,65], the CAPE-II framework provides a comprehensive, standardized scoring rubric that regulators or platform safety teams can use for quality assessment. Given the current regulatory vacuum and documented harms, standardized frameworks such as the CAPE-II are needed. Finally, our results provide actionable targets for improvement, which will be further discussed. Additionally, we provide specific, brief recommendations for various stakeholders to improve the safety of psychotherapy chatbots in Table 2.

Table 2. Recommendations for stakeholders to improve the safety of psychotherapy chatbots.

Stakeholders	Recommendations
Youth	• Understand that chatbots are not therapists • Seek human help for mental health crises • Rely on multiple sources for support, not only GenAI[a] chatbots • Do not share personal or identifying information with chatbots • Report harmful or inappropriate responses using platform feedback mechanisms
Caregivers and parents	• Provide a nonjudgmental space to discuss sensitive topics such as mental health • Ask youth to show you how they use chatbots • Interact with chatbots to improve understanding of their capabilities, limitations, and settings • Understand privacy and data security • Contact policymakers to voice concerns
Platform developers	• Implement evidence-based crisis protocols that include linkage to human-based crisis services • Make privacy policies readable at a sixth-grade level or lower with plain language summaries upfront • Make repeated disclosures throughout conversations that the chatbot is not a human, therapist, or factual database • Verify age and restrict harmful features accordingly (eg, seductive conversations) • Engage diverse domain experts, youth, and families in product development and red teaming
Policymakers and regulators	• Establish independent evaluation and certification for mental health uses, prioritizing safety • Enforce age restrictions • Delineate between entertainment, wellness, and medical device chatbots • Fund AI literacy programs for youth, parents, and educators • Mandate the release of research regarding product harms

Stakeholders	Recommendations
Clinicians	- Stay informed about popular platforms that youth are using - Routinely ask about chatbot use during sessions with youth - Educate youth and caregivers about the benefits and risks of psychotherapy chatbots - Contribute clinical expertise to chatbot development - Document chatbot-related incidents to build an evidence base of harms and benefits
Researchers	- Examine the effectiveness and acceptability of crisis interventions for chatbots - Use youth participatory methods, especially with marginalized groups, to understand real-world usage - Develop standardized evaluation frameworks such as the CAPE-II[b] for quality assessment - Study long-term mental health and development effects of chatbot use through causal or quasi-experimental designs - Create and validate measures of problematic chatbot use in youth

[a]GenAI: generative artificial intelligence.
[b]CAPE-II: Conversational Agent for Psychotherapy Evaluation II framework.

Chatbot Strengths

The strengths of chatbots align with prior qualitative research showing that youth value an accessible and nonjudgmental space to discuss socioemotional challenges [4,6,66]. Free access, whether by computer or mobile device, and availability in multiple languages contribute to their popularity. Still, even if the information is personalized and accurate, youth may have difficulty understanding it, as most chatbot outputs exceed a sixth-grade reading level. In our prior work on OpenAI GPT Store chatbots, we also found poor chatbot output readability [30]. Since readability does not affect information quality [67], chatbot developers should allow users to adjust the reading level. Although we found that chatbots rarely made overtly harmful or unsafe statements, explicitly probing for endorsement or lack of challenging of harmful or delusional statements, as done in other studies, may strengthen this domain. Despite these strengths, given their potential to lead to youth being harmed, 3 lower quality domains warrant further discussion: monitoring and risk evaluation, privacy, and transparency of training data and knowledge base.

Chatbot Quality Concerns

Our findings and others [31,33,68,69] suggest that current GenAI chatbots are not well-equipped for crisis management. They fail to consistently detect crises, and when they do, they often respond in ways that are risky, nonempathetic, and not evidence-based. These failures violate the ethical principle of nonmaleficence. Many did not recognize when depressive and anxiety symptoms were severe enough to require professional support. Prior studies of ChatGPT (OpenAI)-based chatbots found the same [30,33], although Snapchat–My AI, which is ChatGPT-based, always detected a need for escalation of support. It is possible that Snapchat has implemented safeguards to make My AI more sensitive.

Monitoring and Risk Evaluation

Regarding suicidality, the AI Psychologist hosted on CHAI requested a nonevidence-based, no-suicide contract and suggested upgrading to a paid tier to discuss suicidal thoughts more, which is potentially harmful. Replika also suggested upgrading after suicidal thoughts were expressed, though without this condition. There was no direct encouragement of suicide or self-harm as found in other studies of different GenAI chatbots [68,69]. Except for the PolyBuzz–Therapist chatbot, chatbots rarely provided crisis contact information, such as suicide hotlines, unless prompted. Young and colleagues [70] found that young adults preferred AI-generated responses over human responses in multiple domains (relationships, self-expression, and physical health), except in response to suicidal thoughts. Specifically, the youth did not like the immediate recommendation for human support and the chatbot saying that it could not help. Indeed, we found Snapchat–My AI's frequent prompts with mental health resources, even for mild symptoms, off-putting. Thus, a nuanced approach to crisis management may be necessary, balancing chatbot inquisitiveness and acknowledging limitations. However, inviting further discussion of suicidality without human oversight could be problematic given probabilistic outputs, and platforms may avoid it due to liability concerns.

Nevertheless, given the reported harm to youth and lack of regulation, companies must offer solutions. Solutions should balance respect for youth's autonomy with the ethical principles of nonmaleficence and beneficence. Similar to conversations with mental health professionals, chatbots can inform youth at the start of conversations about their limitations and protocols for severe mental health symptoms and suicidality. This gives youth the agency to determine whether they want to share these thoughts with the chatbot. When suicidal thoughts arise, chatbots should promote evidence-based strategies. For example, the chatbot could use a rule-based approach to conduct suicide safety planning by identifying warning signs, coping strategies, and external support systems to address suicidality. This approach may be viable, as there is evidence that a standalone safety plan mobile app, without human support, can reduce suicidality [71]. Sending young people to an external website may raise privacy concerns or leave them feeling abandoned. One solution is to have humans in the loop, such as crisis responders or emergency contacts, who are alerted when youth engage in high-risk behaviors. For example, users could optionally allow caregivers to be alerted when sensitive keywords or language indicating suicidal ideation is detected, as done in a recent study of a GenAI chatbot for depression [72]. However, this approach challenges autonomy, privacy, and scalability. Even if scalable, youth may not want human support or face barriers accessing support. Ultimately, engagement with stakeholders is needed to examine what interventions are effective and feasible.

Privacy

Privacy issues with suicidality speak to larger concerns with DTC GenAI psychotherapy chatbots. Privacy is a major concern for youth and parents [4], though they may not recognize how their privacy is compromised [73]. Our results show that platforms handle privacy poorly. Privacy policies were difficult to understand due to their challenging readability. Another study of Replika and CHAI platforms found the same [74]. Both platforms collect personal data and share app activity; CHAI also shares personal data [74]. The legal and technical jargon in these policies do not align with regulations, such as the General Data Protection Regulation and the California Consumer Privacy Act, that require easy-to-understand language. We advocate for short, plain language or illustrated summaries highlighting the main points of the privacy policy, placed upfront before the full policy to support user comprehension and decision-making [75]. Given the sensitivity of conversational data, future research should explore whether prior recommended summary components (intended purpose of data, collection of data, use of data, and data retention, sharing with third parties) will suffice [76].

We found that platforms provide unclear information about how they use and encrypt user data. Consequently, in the event of a breach or leak of data, others could discover which chatbot a youth used. If the chatbot has “therapist” or “psychologist” in its name, it would undermine the anonymity that many youth seek. Youth may want to remove their data, but the ability to permanently delete personal data is not always specified. Because people may be more likely to share sensitive information with chatbots [77,78], platforms should at least clearly inform users about data collection, sharing, and security so they can decide whether to use a chatbot or share sensitive information. All users should be able to opt out of data collection, especially youth who may not fully understand the consequences of their data being collected and shared.

The tension between privacy and nonmaleficence presents challenges for psychotherapy chatbots. Platforms collect and store conversational data for multiple purposes, including legal reasons or product improvement. Storage of high-risk conversations may be warranted. When long-term storage of high-risk conversations is unavoidable, there should be end-to-end encryption with access limited to authorized stakeholders (eg, safety teams). In addition to these measures, this high-risk data should not be used for model training to minimize risks to privacy.

Transparency

Expanding upon transparency, across platforms, we were unable to find information on chatbots’ training data and knowledge base. Our results expand upon prior studies of ChatGPT [30,58] where this information was unavailable. This opacity is problematic for multiple reasons. First, transparency is needed to determine trustworthiness [79]. A recent study found that 50% of adolescents distrust the information or advice that chatbots provide [23]. However, trust was higher among younger adolescents [23], who also have a more positive impression of chatbots [38], suggesting they may be more susceptible to believing chatbot outputs. Providing information on the training data and knowledge base is one way to help youth, parents, and other stakeholders determine the trustworthiness of model outputs, which ethically supports autonomy through informed consent. In addition, transparency on how that information is used to generate outputs (ie, explainable AI) is also needed. This could be accomplished through citations with reasoning traces showing the steps used to arrive at a response. Like outputs, these explanations should be at a reading level appropriate for users to aid understanding.

Second, transparency is needed to support the ethical principle of justice because it is unclear if models are representative of diverse experiences. GenAI chatbots may propagate biases and output insensitive responses to marginalized and underrepresented populations. For example, young adults in India and LGBTQ+ (Lesbian, Gay, Bisexual, Transgender, Queer, Questioning Plus) participants reported culturally incongruent or harmful advice [6,80,81], such as quitting a job they financially rely on due to workplace homophobia or recommending interventions that do not align with non-Western family norms. Platforms should disclose training data and knowledge base details to ensure diverse perspectives and lived experiences are included. Furthermore, to mitigate these potential harms, youth from marginalized groups should be represented in co-design and advisory boards [82,83]. Also, there should be a feedback channel that lets young users flag biased outputs to promote change.

Mitigating Risks

More broadly, we believe red teaming, proactively testing for vulnerabilities, can help mitigate risks. In this approach, GenAI chatbot developers work with stakeholders who have relevant professional expertise or lived experience to test their chatbots. While we conducted our evaluation independently, our approach and framework can be applied in red teaming. With a clear set of criteria, our approach can help standardize these efforts. Given the rapid development of LLMs and changes to DTC chatbots, evaluations of chatbots should occur each time there is a technology update or, at minimum, semiannually, as updates may not be announced.

Still, our approach may not address the potential harms to raters conducting this work. Ideally, red teaming would include members of vulnerable populations, especially youth from historically marginalized groups, since their interactions with chatbots may differ from those of mental health professionals. However, involving vulnerable populations requires careful adherence to guidelines that minimize harm risks and provide comprehensive support [84]. Financial compensation alone is not adequate protection for participants who may be harmed during testing. Psychological support should be offered. While not mutually exclusive, a less direct approach would be to create a youth advisory board that could review issues identified during red teaming and suggest other areas to explore. As an alternative, some have used LLMs to simulate humans to rate quality. While LLMs could potentially simulate users at scale, human simulation studies have found that LLM responses are less varied, more socially desirable, and biased [85-87]. Future research should explore whether simulating a human persona is a valid way to evaluate GenAI psychotherapy chatbots.

Limitations and Future Directions

This study has limitations. In total, 3 of the platforms we selected, Character.AI, PolyBuzz, and CHAI offer users a wide selection of chatbots with preset parameters or the option to create a custom bot. Due to resource constraints, we were unable to evaluate multiple chatbots on each platform and instead focused on the most popular psychotherapy chatbot. Also, it is possible that additional personas could further elicit information, but given resource constraints, this was not possible. Although in our prior study and this study, we found no differences in performance between the 2 personas used. An additional limitation is that our interactions with the chatbots were primarily conducted within 1 interaction. However, users may interact with chatbots multiple times during the course of a conversation. To address this limitation, we interacted with chatbots 24 hours after the initial conversation to see if information was retained. Future research should explore longer-term interactions. Furthermore, all chatbots evaluated were hosted on platforms that portray them as entertainment or wellness products, rather than as mental health support. Despite this, users may turn to these chatbots for therapeutic relief, and several chatbots are described as supporting mental health, so assessing their psychotherapeutic capabilities is essential. Future research should compare the quality of evidence- and rule-based chatbot (eg, Wysa; Touchkin eServices Private Limited) or GenAI chatbot (eg, Therabot; Center for Technology and Behavioral Health, Dartmouth College) with these widely used entertainment and wellness GenAI chatbots.

Our goal was to develop a reproducible framework for efficiently evaluating the benefits and challenges of psychotherapy chatbots for youth. This framework was informed by the input of youth who were part of the evaluation team. Nonetheless, we recognize that a youth participatory research approach, as well as narrative and phenomenological qualitative approaches with youth, may further advance framework development. Having younger youth engage with chatbots would be particularly beneficial, given that our team consists of youth in their early 20s. Evaluation instructions may need to be adapted for younger youth to use.

Findings from our evaluation can support immediate fixes such as crisis support, but they do not address the underlying reasons why youth use chatbots for mental health support. Loneliness, lack of social connection, barriers to basic needs, and other social and structural determinants of health all contribute to youth turning to GenAI chatbots rather than human support. This reality makes evaluation more pressing as chatbots may be the only resource available to youth, especially those from marginalized communities. Longer-term psychosocial and structural solutions are needed to support youth.

Conclusions

These findings reveal that DTC GenAI psychotherapy chatbots are unsafe for the millions of youth who use them. While they have strengths that support beneficence and justice through accessibility, they create unacceptable risks: improper handling of crises violates nonmaleficence, and a lack of transparency regarding privacy and model training undermines autonomy. We advocate for immediate reforms to protect this vulnerable population. The CAPE-II framework and other comprehensive chatbot evaluation frameworks are valuable tools to identify and inform stakeholders, particularly platforms and policymakers, where reforms are needed.

Acknowledgments

Generative artificial intelligence was used to support reference formatting.

Funding

This work is supported by the University of California, Los Angeles, Depression Grand Challenges Judith Young Fellowship awarded to KS. The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the article; and decision to submit the article for publication.

Data Availability

Data collected and used in this study are available upon reasonable request from the corresponding author.

Authors' Contributions

KS conceived the study and provided project administration, resources, supervision, and software. All authors contributed to methodology, data curation, and investigation. KS conducted the formal analysis.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Additional methods and results.

[DOCX File (Microsoft Word File), 59 KB-Multimedia Appendix 1]

Checklist 1

STROBE checklist.

[PDF File (Adobe File), 192 KB-Checklist 1]

References

1. Yu Y, Liu Y, Zhang J, Huang Y, Wang Y. Understanding generative AI risks for youth: a taxonomy based on empirical data. arXiv. Preprint posted online on Feb 22, 2025. [doi: 10.48550/arXiv.2502.16383]
2. Brandtzæg PB, Skjuve M, Følstad A. Emerging AI-individualism: how young people integrate social AI into their lives. SSRN Journal. Jun 2024. [doi: 10.2139/ssrn.4836120]
3. Online Nation 2024 report. Ofcom; 2024. URL: https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-nation/2024/online-nation-2024-report.pdf [Accessed 2024-12-10]
4. Yu Y, Sharma T, Hu M, Wang J, Wang Y. Exploring parent-child perceptions on safety in generative AI: concerns, mitigation strategies, and design implications. Presented at: 2025 IEEE Symposium on Security and Privacy (SP); May 12-15, 2025; San Francisco, United States. [doi: 10.1109/SP61157.2025.00090]
5. Ta-Johnson VP, Boatfield C, Wang X, et al. Assessing the topics and motivating factors behind human-social chatbot interactions: thematic analysis of user experiences. JMIR Hum Factors. Oct 3, 2022;9(4):e38876. [doi: 10.2196/38876] [Medline: 36190745]
6. Kavitha K, Joshith VP, Sharma S. Beyond text: ChatGPT as an emotional resilience support tool for Gen Z – a sequential explanatory design exploration. eLDM. Jun 3, 2024. [doi: 10.1177/20427530241259099]
7. Maples B, Cerit M, Vishwanath A, Pea R. Loneliness and suicide mitigation for students using GPT3-enabled chatbots. Npj Ment Health Res. Jan 22, 2024;3(1):4. [doi: 10.1038/s44184-023-00047-6] [Medline: 38609517]
8. Herbener AB, Damholdt MF. Are lonely youngsters turning to chatbots for companionship? The relationship between chatbot usage and social connectedness in Danish high-school students. Int J Hum Comput Stud. Feb 2025;196:103409. [doi: 10.1016/j.ijhcs.2024.103409]
9. McGorry PD, Mei C, Dalal N, et al. The Lancet Psychiatry Commission on youth mental health. Lancet Psychiatry. Sep 2024;11(9):731-774. [doi: 10.1016/S2215-0366(24)00163-9] [Medline: 39147461]
10. Madden M, Calvin A, Hasse A, Lenhart A. The Dawn of the AI Era: Teens, Parents, and the Adoption of Generative AI at Home and School. Common Sense; 2024. URL: https://www.commonsensemedia.org/sites/default/files/research/report/2024-the-dawn-of-the-ai-era_final-release-for-web.pdf [Accessed 2025-11-11]
11. Dahl RE, Allen NB, Wilbrecht L, Suleiman AB. Importance of investing in adolescence from a developmental science perspective. Nature New Biol. Feb 21, 2018;554(7693):441-450. [doi: 10.1038/nature25770] [Medline: 29469094]
12. Bansal B. UK trials AI therapy: young adults most open to talking to AI-powered chatbots about mental health. YouGov. 2024. URL: https://business.yougov.com/content/49481-uk-trials-ai-therapy [Accessed 2025-09-05]
13. Hoffman BD, Oppert ML, Owen M. Understanding young adults' attitudes towards using AI chatbots for psychotherapy: the role of self-stigma. Comput Hum Behav Artif Hum. Aug 2024;2(2):100086. [doi: 10.1016/j.chbah.2024.100086]
14. Gulliver A, Griffiths KM, Christensen H. Perceived barriers and facilitators to mental health help-seeking in young people: a systematic review. BMC Psychiatry. Dec 30, 2010;10(1):113. [doi: 10.1186/1471-244X-10-113] [Medline: 21192795]
15. Radez J, Reardon T, Creswell C, Lawrence PJ, Evdoka-Burton G, Waite P. Why do children and adolescents (not) seek and access professional help for their mental health problems? A systematic review of quantitative and qualitative studies. Eur Child Adolesc Psychiatry. Feb 2021;30(2):183-211. [doi: 10.1007/s00787-019-01469-4] [Medline: 31965309]
16. Heinz MV, Mackin DM, Trudeau BM, et al. Randomized trial of a generative AI chatbot for mental health treatment. NEJM AI. Mar 27, 2025;2(4). [doi: 10.1056/AIoa2400802]
17. Campellone TR, Flom M, Montgomery RM, et al. Safety and user experience of a generative artificial intelligence digital mental health intervention: exploratory randomized controlled trial. J Med Internet Res. May 23, 2025;27:e67365. [doi: 10.2196/67365] [Medline: 40408143]
18. Parks A, Travers E, Perera-Delcourt R, Major M, Economides M, Mullan P. Is this chatbot safe and evidence-based? A call for the critical evaluation of generative AI mental health chatbots. J Particip Med. May 29, 2025;17:e69534. [doi: 10.2196/69534] [Medline: 40440646]

19. Nasr M, Carlini N, Hayase J, et al. Scalable extraction of training data from (production) language models. arXiv. Preprint posted online on Nov 28, 2023. [doi: 10.48550/arXiv.2311.17035]
20. Carlini N, Tramèr F, Wallace E, et al. Extracting training data from large language models. Presented at: Proceedings of the 30th USENIX Security Symposium; Aug 11-13, 2021; Berkeley, CA, USA. URL: https://www.usenix.org/system/files/sec21-carlini-extracting.pdf [Accessed 2025-11-11]
21. Wells K. An eating disorders chatbot offered dieting advice, raising fears about AI in health. NPR. Jun 8, 2023. URL: https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea [Accessed 2024-11-27]
22. Zhang Z, Shen C, Yao B, Wang D, Li T. Secret use of large language model (LLM). Proc ACM Hum-Comput Interact. May 2, 2025;9(2):1-26. [doi: 10.1145/3711061]
23. Robb MB, Mann S. Talk, trust, and trade-offs: how and why teens use AI companions. Common Sense Media; Jul 16, 2025. URL: https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions [Accessed 2025-09-06]
24. Eira M, Rasouli A, Charisi V. Parents' perceptions about the use of generative AI systems by adolescents. Presented at: IDC '25; Jun 23-26, 2025; Reykjavik, Iceland. [doi: 10.1145/3713043.3731508]
25. Brittain B. Google, AI firm must face lawsuit filed by a mother over suicide of son, US court says. Reuters. May 21, 2025. URL: https://www.reuters.com/sustainability/boards-policy-regulation/google-ai-firm-must-face-lawsuit-filed-by-mother-over-suicide-son-us-court-says-2025-05-21/ [Accessed 2025-09-05]
26. Hill K. A teen was suicidal. ChatGPT was the friend he confided in. The New York Times. 2025. URL: https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html [Accessed 2025-09-05]
27. Letter to the FTC on the risks of underregulated generative AI and its impact on vulnerable populations. American Psychological Association. APA Services; 2024. URL: https://www.apaservices.org/advocacy/generative-ai-technology-regulation-concern.pdf [Accessed 2025-12-02]
28. Bansal B. Can an AI chatbot be your therapist? A third of Americans are comfortable with the idea. YouGov. May 18, 2024. URL: https://yougov.com/en-us/articles/49480-can-an-ai-chatbot-be-your-therapist [Accessed 2025-09-05]
29. Bond BJ, Parent MC, Willie L, Green AE. Parasocial relationships, AI chatbots, and joyful online interactions among a diverse sample of LGBTQ+ young people. Hopelab; Sep 30, 2024. URL: https://assets.hopelab.org/wp-content/uploads/2024/09/HL_ParasocialRelationships_report_FINALpdf.pdf [Accessed 2025-11-11]
30. Sobowale K, Humphrey DK. Evaluating the quality of psychotherapy conversational agents: framework development and cross-sectional study. JMIR Form Res. Jul 2, 2025;9:e65605. [doi: 10.2196/65605] [Medline: 40600851]
31. Moore J, Grabb D, Agnew W, et al. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. Presented at: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency; Jun 23-26, 2025; Athens, Greece. [doi: 10.1145/3715275.3732039]
32. Clark A. The ability of AI therapy bots to set limits with distressed adolescents: simulation-based comparison study. JMIR Ment Health. Aug 18, 2025;12:e78414. [doi: 10.2196/78414] [Medline: 40825182]
33. Heston TF. Evaluating risk progression in mental health chatbots using escalating prompts. medRxiv. Preprint posted online on Sep 12, 2023. [doi: 10.1101/2023.09.10.23295321]
34. Moore O. The top 100 gen AI consumer apps. Andreessen Horowitz. URL: https://a16z.com/100-gen-ai-apps [Accessed 2024-06-07]
35. Laestadius L, Bishop A, Gonzalez M, Illenčík D, Campos-Castillo C. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. New Media Soc. Oct 2024;26(10):5923-5941. [doi: 10.1177/14614448221142007]
36. Faverio M, Sidoti O. Teens, social media and technology 2024. Pew Research Center; Dec 12, 2024. URL: https://www.pewresearch.org/internet/2024/12/12/teens-social-media-and-technology-2024/ [Accessed 2025-09-06]
37. Snap Inc. 2023 investor day – recap. Snap Newsroom. 2023. URL: https://newsroom.snap.com/investor-day-2023 [Accessed 2025-09-06]
38. Vanhoffelen G, Vandenbosch L, Schreurs L. Teens, tech, and talk: adolescents' use of and emotional reactions to Snapchat's my AI chatbot. Behav Sci (Basel). Jul 30, 2025;15(8):1037. [doi: 10.3390/bs15081037] [Medline: 40867394]
39. Q1 2025 investor letter. Snap Inc; 2025. URL: https://investor.snap.com/files/doc_financials/2025/q1/Q1-2025-Investor-Letter.pdf [Accessed 2025-09-06]
40. Dogruel L, Joeckel S, Bowman ND. Choosing the right app: an exploratory perspective on heuristic decision processes for smartphone app selection. Mob Media Commun. Jan 2015;3(1):125-144. [doi: 10.1177/2050157914557509]
41. Yuguero O, Ruiz-Trujillo P, Esquerda M, Terribas N, Aymerich M. Applying the principle of justice in digital health. NPJ Digit Med. Jul 21, 2025;8(1):467. [doi: 10.1038/s41746-025-01877-8] [Medline: 40691313]

42. Sim I, Cassel C. The ethics of relational AI - expanding and implementing the Belmont principles. N Engl J Med. Jul 18, 2024;391(3):193-196. [doi: 10.1056/NEJMp2314771] [Medline: 39007542]
43. Silva GRS, Canedo ED. Towards user-centric guidelines for chatbot conversational design. Int J Hum Comput Interact. Jan 17, 2024;40(2):98-120. [doi: 10.1080/10447318.2022.2118244]
44. Torous JB, Chan SR, Gipson S, et al. A hierarchical framework for evaluation and informed decision making regarding smartphone apps for clinical care. Psychiatr Serv. May 1, 2018;69(5):498-500. [doi: 10.1176/appi.ps.201700423] [Medline: 29446337]
45. Lee YK, Lee I, Shin M, Bae S, Hahn S. Chain of empathy: enhancing empathetic response of large language models based on psychotherapy models. arXiv. Preprint posted online on Nov 2, 2023. [doi: 10.48550/arXiv.2311.04915]
46. Li E, Kealy D, Aafjes-van Doorn K, et al. “It felt like I was being tailored to the treatment rather than the treatment being tailored to me”: Patient experiences of helpful and unhelpful psychotherapy. Psychother Res. Jun 2025;35(5):695-709. [doi: 10.1080/10503307.2024.2360448] [Medline: 38833539]
47. Ethical principles of psychologists and code of conduct. American Psychological Association. 2017. URL: https://www.apa.org/ethics/code [Accessed 2025-11-11]
48. Chaszczewicz A, Shah R, Louie R, Arnow B, Kraut R, Yang D. Multi-level feedback generation with large language models for empowering novice peer counselors. Presented at: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Aug 11-16, 2024; Bangkok, Thailand. [doi: 10.18653/v1/2024.acl-long.227]
49. Liu S, Zheng C, Demasi O, et al. Towards emotional support dialog systems. Presented at: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); Aug 1-6, 2021. [doi: 10.18653/v1/2021.acl-long.269]
50. Wampold BE, Flückiger C. The alliance in mental health care: conceptualization, evidence and clinical applications. World Psychiatry. Feb 2023;22(1):25-41. [doi: 10.1002/wps.21035] [Medline: 36640398]
51. Meng J, Rheu M (MJ), Zhang Y, Dai Y, Peng W. Mediated social support for distress reduction: AI chatbots vs. human. Proc ACM Hum-Comput Interact. Apr 14, 2023;7(CSCW1):1-25. [doi: 10.1145/3579505]
52. Rheu M (MJ), Dai Y (N), Meng J, Peng W. When a chatbot disappoints you: expectancy violation in human-chatbot interaction in a social support context. Communic Res. Oct 2024;51(7):782-814. [doi: 10.1177/00936502231221669]
53. Boswell JF, Hepner KA, Lysell K, et al. The need for a measurement-based care professional practice guideline. Psychotherapy (Chic). Mar 2023;60(1):1-16. [doi: 10.1037/pst0000439] [Medline: 35771518]
54. Coghlan S, Leins K, Sheldrick S, Cheong M, Gooding P, D’Alfonso S. To chat or bot to chat: ethical issues with using chatbots in mental health. Digit Health. 2023;9:20552076231183542. [doi: 10.1177/20552076231183542] [Medline: 37377565]
55. About our methodology. Mozilla Foundation. URL: https://foundation.mozilla.org/en/privacynotincluded/about/methodology/#minimum-security-standards [Accessed 2024-08-16]
56. Zhan H, Zheng A, Lee YK, Suh J, Li JJ, Ong DC. Large language models are capable of offering cognitive reappraisal, if guided. arXiv. Preprint posted online on Apr 1, 2024. [doi: 10.48550/arXiv.2404.01288]
57. Ramos G, Ponting C, Labao JP, Sobowale K. Considerations of diversity, equity, and inclusion in mental health apps: a scoping review of evaluation frameworks. Behav Res Ther. Dec 2021;147:103990. [doi: 10.1016/j.brat.2021.103990] [Medline: 34715396]
58. Golden A, Aboujaoude E. The framework for AI tool assessment in mental health (FAITA - Mental Health): a scale for evaluating AI-powered mental health tools. World Psychiatry. Oct 2024;23(3):444-445. [doi: 10.1002/wps.21248] [Medline: 39279357]
59. Kieling C, Buchweitz C, Caye A, et al. Worldwide prevalence and disability from mental disorders across childhood and adolescence: evidence from the global burden of disease study. JAMA Psychiatry. Apr 1, 2024;81(4):347-356. [doi: 10.1001/jamapsychiatry.2023.5051] [Medline: 38294785]
60. Hall EB, Mufson L. Interpersonal psychotherapy for depressed adolescents (IPT-A): a case illustration. J Clin Child Adolesc Psychol. Jul 2009;38(4):582-593. [doi: 10.1080/15374410902976338] [Medline: 20183644]
61. Jiménez Chafey MI, Bernal G, Rosselló J. Clinical case study: CBT for depression in a Puerto Rican adolescent: challenges and variability in treatment response. Depress Anxiety. 2009;26(1):98-103. [doi: 10.1002/da.20457] [Medline: 18781640]
62. Young JE, Rygh JL, Weinberger AD, Beck AT. Cognitive therapy for depression. In: Clinical Handbook of Psychological Disorders: A Step-by-Step Treatment Manual. Guilford Press; 2008:250-305.
63. Krippendorff K. Reliability in content analysis: some common misconceptions and recommendations. Human Comm Res. Jul 2004;30(3):411-433. [doi: 10.1111/j.1468-2958.2004.tb00738.x]

64. Sandelowski M. Whatever happened to qualitative description? Res Nurs Health. Aug 2000;23(4):334-340. [Medline: 10940958]
65. Chandra M, Naik S, Ford D, et al. From lived experience to insight: unpacking the psychological risks of using ai conversational agents. Presented at: FAccT '25 Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency; Jun 23-26, 2025; Athens, Greece. [doi: 10.1145/3715275.3732063]
66. Ma Z, Mei Y, Su Z. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. arXiv. Preprint posted online on Jul 28, 2023. [doi: 10.48550/arXiv.2307.15810]
67. Musheyev D, Pan A, Gross P, et al. Readability and information quality in cancer information from a free vs paid chatbot. JAMA Netw Open. Jul 1, 2024;7(7):e2422275. [doi: 10.1001/jamanetworkopen.2024.22275] [Medline: 39058491]
68. De Freitas J, Cohen IG. The health risks of generative AI-based wellness apps. Nat Med. May 2024;30(5):1269-1275. [doi: 10.1038/s41591-024-02943-6] [Medline: 38684859]
69. Grabb D, Lamparth M, Vasan N. Risks from language models for automated mental healthcare: ethics and structure for implementation. Presented at: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; Oct 20-22, 2025; Madrid, Spain. [doi: 10.5555/3716662.3716705]
70. Young J, Jawara LM, Nguyen DN, Daly B, Huh-Yoo J, Razi A. The role of AI in peer support for young people: a study of preferences for human- and AI-generated responses. Presented at: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu, HI, USA. [doi: 10.1145/3613904.3642574]
71. Rainbow C, Tatnell R, Blashki G, Fuller-Tyszkiewicz M, Melvin GA. Digital safety plan effectiveness and use: findings from a three-month longitudinal study. Psychiatry Res. Mar 2024;333:115748. [doi: 10.1016/j.psychres.2024.115748] [Medline: 38277811]
72. Li Y, Ding X, Chen Y, Li Y, Ma N. Customizable AI for depression care: improving the user experience of large language model-driven chatbots. Presented at: Proceedings of the 2025 ACM Designing Interactive Systems Conference; Jul 5-9, 2025; Madeira, Portugal. [doi: 10.1145/3715336.3735795]
73. Bae Brandtzæg PB, Skjuve M, Kristoffer Dysthe KK, Følstad A. When the social becomes non-human: young people's perception of social support in chatbots. Presented at: CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan. [doi: 10.1145/3411764.3445318]
74. Ragab A, Mannan M, Youssef A. "Trust Me Over My Privacy Policy": Privacy Discrepancies in Romantic AI Chatbot Apps. Presented at: 2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW); Jul 8-12, 2024; Vienna, Austria. [doi: 10.1109/EuroSPW61312.2024.00060]
75. Alfawzan N, Christen M, Spitale G, Biller-Andorno N. Privacy, data sharing, and data security policies of women's mHealth apps: scoping review and content analysis. JMIR Mhealth Uhealth. May 6, 2022;10(5):e33735. [doi: 10.2196/33735] [Medline: 35522465]
76. Tomuro N, Lytinen S, Hornsburg K. Automatic summarization of privacy policies using ensemble learning. Presented at: CODASPY '16: Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy; Mar 9-11, 2016; New Orleans, LA, USA. [doi: 10.1145/2857705.2857741]
77. Gieselmann M, Sassenberg K. The more competent, the better? The effects of perceived competencies on disclosure towards conversational artificial intelligence. Soc Sci Comput Rev. Dec 2023;41(6):2342-2363. [doi: 10.1177/08944393221142787]
78. Atleson M. The luring test: AI and the engineering of consumer trust. New York University School of Law. May 1, 2023. URL: https://wp.nyu.edu/compliance_enforcement/2024/04/17/the-luring-test-ai-and-the-engineering-of-consumer-rust/ [Accessed 2024-12-05]
79. Kovari A. Explainable AI chatbots towards XAI ChatGPT: a review. Heliyon. Jan 30, 2025;11(2):e42077. [doi: 10.1016/j.heliyon.2025.e42077] [Medline: 39906828]
80. Ma Z, Mei Y, Long Y, Su Z, Gajos KZ. Evaluating the experience of LGBTQ+ people using large language model based chatbots for mental health support. Presented at: CHI '24: Proceedings of the CHI Conference on Human Factors in Computing Systems; May 11-16, 2024; Honolulu, HI, USA. [doi: 10.1145/3613904.3642482]
81. Song I, Pendse SR, Kumar N, De Choudhury M. The typing cure: experiences with large language model chatbots for mental health support. Presented at: Proceedings of the ACM on Human-Computer Interaction; Oct 18-22, 2025; Bergen, Norway. [doi: 10.1145/3757430]
82. Poulsen A, Hickie IB, de Haan Z, et al. Co-designing a conversational generative artificial intelligence system for youth mental health. Presented at: CHI EA '25: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems; May 11-16, 2025; Yokohama, Japan. [doi: 10.1145/3706599.3719696]
83. Figueroa CA, Ramos G, Psihogios AM, et al. Advancing youth co-design of ethical guidelines for AI-powered digital mental health tools. Nat Mental Health. Jul 31, 2025;3(8):870-878. [doi: 10.1038/s44220-025-00467-7]

84. Pendse SR, Gergle D, Kornfield R, et al. When testing AI tests us: safeguarding mental health on the digital frontlines. Presented at: FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency; Jun 23-26, 2025; Athens, Greece. [doi: 10.1145/3715275.3732120]
85. Wang A, Morgenstern J, Dickerson JP. Large language models that replace human participants can harmfully misportray and flatten identity groups. Nat Mach Intell. 2025;7(3):400-411. [doi: 10.1038/s42256-025-00986-z] [Medline: 40416329]
86. Gao Y, Lee D, Burtch G, Fazelpour S. Take caution in using LLMs as human surrogates. Proc Natl Acad Sci U S A. Jun 17, 2025;122(24):e2501660122. [doi: 10.1073/pnas.2501660122] [Medline: 40512797]
87. Lin Z. Six fallacies in substituting large language models for human participants. Adv Methods Pract Psychol Sci. Jul 2025;8(3):1-19. [doi: 10.1177/25152459251357566]

Abbreviations

CAPE-II: Conversational Agent for Psychotherapy Evaluation II framework
DTC: direct-to-consumer
GenAI: generative artificial intelligence
IRR: interrater reliability
LLM: large language model
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

Edited by John Torous; peer-reviewed by Bryanna Moore, Ian Steenstra, Joel Wester; submitted 29.Jun.2025; final revised version received 28.Sep.2025; accepted 29.Sep.2025; published 10.Dec.2025

Please cite as:
Sobowale K, Humphrey DK, Zhao SY
Evaluating Generative AI Psychotherapy Chatbots Used by Youth: Cross-Sectional Study
JMIR Ment Health 2025;12:e79838
URL: https://mental.jmir.org/2025/1/e79838
doi: 10.2196/79838