frontiers in Digital Health

TYPE Original Research
PUBLISHED 10 June 2026
DOI 10.3389/fdgth.2026.1797681

Check for updates

OPEN ACCESS

EDITED BY
Michael Patrick Schaub,
University of Zurich, Switzerland

REVIEWED BY
Warren Mansell,
The University of Manchester, United Kingdom
Marissa De Vries,
Arkin, Netherlands

*CORRESPONDENCE
Xintong Lu
xintong.lu@parkview.com

RECEIVED 27 January 2026
REVISED 09 April 2026
ACCEPTED 16 April 2026
PUBLISHED 10 June 2026

CITATION
Lu X, Guha S, Pfafman R, Pater JA, Washington CL and Nova FF (2026) Trauma-informed conversational agents for mental health: understanding user perspectives and experiences. Front. Digit. Health 8:1797681. doi: 10.3389/fdgth.2026.1797681

COPYRIGHT
© 2026 Lu, Guha, Pfafman, Pater, Washington and Nova. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Trauma-informed conversational agents for mental health: understanding user perspectives and experiences

Xintong Lu[1]*, Shion Guha[1,2,3], Rachel Pfafman[1], Jessica A. Pater[1], Courtney L. Washington[4,5] and Fayika Farhat Nova[1]

[1]Parkview Health, Parkview Research Center, Fort Wayne, IN, United States, [2]Department of Computer Science, University of Toronto, Toronto, ON, Canada, [3]Faculty of Information, University of Toronto, Toronto, ON, Canada, [4]Parkview Health, Parkview Behavioral Health Institute, Fort Wayne, IN, United States, [5]Psychological Science Department, Northern Michigan University, Marquette, MI, United States

Introduction: Mental health conversational agents (MHCAs) offer scalable, accessible psychological support yet raise concerns about safety and appropriateness for trauma, exposed users. While trauma-informed care (TIC) principles, emphasizing safety, trust, empowerment, collaboration, peer support, and cultural sensitivity, are well-established in clinical practice, their application and user interpretation in chatbot contexts remain unexplored. This study aimed to explore (1) how users conceptualize trauma-informed care in the context of mental health chatbots and (2) what factors can help us predict whether users perceive their chatbot interaction as trauma-informed.

Methods: A 59-item web-based survey (REDCap) assessed demographics, technology proficiency, adverse life experiences, chatbot usage, and TIC perceptions among 606 participants recruited via ClickWorker, social media, and MyChart. Exploratory and confirmatory factor analyses identified latent TIC domains. Multivariable logistic regression determined predictors of trauma-informed perception.

Results: High satisfaction (84.5%) and trauma-informed perception (92.9%) were reported. Factor analyses validated a five-domain structure: Trust & Transparency, Data Safety, Empowerment, Peer Support, and Cultural Sensitivity (CFI = 0.978; RMSEA = 0.045). Trust (OR = 3.89, p = .001), Empowerment (OR = 1.97, p = .025), and Peer Support (OR = 1.73, p = .021) significantly predicted trauma-informed perception. Convenience-driven use (OR = 0.33, p = .041) and smartphone proficiency (OR = 0.38, p = .033) showed negative associations.

Discussion: This study provides empirical foundation for measuring trauma-informed design in MHCAs. Trust-building, emotional validation, and peer connection significantly shape user perceptions. Findings emphasize the necessity of aligning AI-based mental health tools with trauma-informed principles to enhance therapeutic safety, credibility, and inclusivity across diverse user populations.

KEYWORDS
artificial intelligence, chatbots, conversational agents, human-AI interaction, mental health, trauma-informed care

1 Introduction

Conversational agents (CAs), or chatbots, are AI-driven systems designed to simulate natural human conversation through text or voice interfaces [1]. In mental health contexts, these technologies are increasingly used to deliver psychoeducation, self-guided therapy modules, and supportive dialogue that mimics elements of traditional psychotherapy [2, 3]. Several studies suggest that mental health conversational agents (MHCAs) can improve emotional well-being, reduce symptoms of depression and anxiety, and enhance users' self-efficacy and motivation to seek care [4–7]. These tools typically draw on established therapeutic models such as cognitive behavioral therapy (CBT), mindfulness, and positive psychology, and provide continuous access, anonymity, and convenience, which may be particularly attractive for individuals with unmet mental health needs [8–10]. As a result, commercial and research-based MHCAs such as Woebot, Wysa, and Youper have proliferated across digital marketplaces and health systems [8, 9, 11, 12].

Despite their growing popularity, MHCAs raise questions about safety, ethical design, and therapeutic efficacy, particularly for vulnerable users [13–16]. Reviews have noted that the evidence base remains mixed, with heterogeneous outcomes and high risk of bias in many studies [4, 10]. Mental health chatbots span a heterogeneous landscape of systems, from earlier rule-based and scripted conversational agents to contemporary large language model (LLM)-based generative AI systems such as those underpinning Woebot and Wysa [17, 18]. While earlier systems relied on predefined scripts and limited natural language understanding, generative AI systems introduce distinct challenges, including the potential for plausible but clinically inappropriate responses, inconsistent emotional attunement, and limited accountability [19, 20]. Empirical comparisons of LLM-generated responses with those of licensed therapists have found that chatbots tend toward directive, advice-giving communication rather than the Socratic, exploratory dialogue characteristic of effective psychotherapy, and are currently considered unsuitable as standalone therapeutic agents [20]. A scoping review of 101 studies further identified privacy and confidentiality (61% of articles) and safety and harm (52%) as the most frequently discussed ethical concerns in conversational AI mental health research [16].High-profile incidents, including reports of inappropriate responses to suicidal ideation or encouragement of harmful behaviors, have heightened public concern and exposed critical limitations in design and oversight across both rule-based and AI-generated conversational systems [21–23].

Trauma-informed care (TIC) has emerged as a foundational framework for addressing safety and ethical considerations in both clinical and digital mental health contexts. Originally proposed by the Substance Abuse and Mental Health Services Administration (SAMHSA), TIC is a strengths-based approach to service delivery that emphasizes understanding and responding to the widespread, disempowering effects of trauma [24–26]. Rather than functioning as a discrete treatment modality, TIC involves acknowledging the impact of trauma and intentionally responding in ways that support safety and avoid retraumatization [27]. The framework is structured around six core principles: (1) safety; (2) trustworthiness and transparency; (3) peer support; (4) collaboration and mutuality; (5) empowerment, voice, and choice; and (6) attention to cultural, historical, and gender issues [27]. Given that many individuals experience traumatic events throughout their lives, regardless of their diagnoses or presenting conditions, TIC has become a foundational approach in mental health care and has been extended beyond traditional clinical settings to domains such as trauma-informed telehealth and trauma-informed computing [28–34]. Critically, recent scholarship has raised concern that generative AI systems may actively increase the risk of technology-mediated trauma among vulnerable populations, underscoring the urgency of applying trauma-informed principles to AI-enabled health tools [28]. A recent scoping review of 118 studies systematically mapping TIC principles in digital technology design found that while safety, trustworthiness, and transparency were the most commonly referenced principles, a majority of studies invoked trauma-informed language without adherence to standardized frameworks, revealing a persistent gap between TIC aspiration and implementation in digital contexts [24]. These trauma-informed approaches aim to maximize physical, psychological, and emotional safety in all health care interactions, fostering opportunities for empowerment, control, and healing through safe, collaborative patient-clinician relationships [31, 34].

Recently, scholars in human-computer interaction and digital health have advocated integrating TIC into the design of digital platforms and AI-enabled technologies, giving rise to the concept of trauma-informed computing [29, 35–37]. Parallel developments in companion robot design have similarly demonstrated the feasibility of embedding TIC principles, including safety, trust, self-compassion, and self-efficacy, into AI system design for trauma-affected populations, providing an important proof of concept for trauma-informed AI development more broadly [25].A trauma-informed chatbot would not only protect user privacy and emotional safety, but also communicate its capabilities and limitations clearly, support user autonomy over how and when to engage, and avoid language or interaction patterns that may be experienced as coercive, dismissive, or triggering [29]. Despite growing advocacy for trauma-informed design, a critical gap remains: existing research has primarily approached TIC from a design or developer perspective, with limited empirical attention to end-user experiences [24, 33, 38]. In particular, there is no well-established, user-centered evidence base that systematically examines how individuals perceive and respond to TIC principles in MHCA interactions, nor which design elements most strongly influence perceptions of psychological safety, trust, and empowerment in automated contexts. The present study directly addresses this gap by empirically investigating end-user perceptions of trauma-informed care in MHCA interactions. Specifically, we examine:

1. **RQ1:** How do users conceptualize TIC principles in the context of mental health chatbots?
2. **RQ2:** What factors predict whether participants perceive their chatbot experience as trauma-informed?

By centering user experiences and expectations, this study aims to clarify what trauma-informed care looks like in MHCA

interactions and to identify actionable features that support perceived emotional safety, trust, and autonomy. The findings provide insights for developers, clinicians, and researchers seeking to implement TIC in digital mental health tools and ensure that MHCAs serve users in an ethical, supportive, and trauma-sensitive manner.

2 Materials and methods

2.1 Survey design

This study was conducted as part of a larger federally funded research project and was approved by the institutional review board (IRB) of the Parkview Health. A structured, self-administered online questionnaire was developed and administered via REDCap to assess users' experiences with MHCAs, their perceptions of trauma-informed chatbot characteristics, and their overall satisfaction with chatbot interactions.

Given the lack of established instruments that operationalize trauma-informed care principles through end-user perspectives specifically in the context of human–AI chatbot interactions, we developed an instrument by mapping SAMHSA's TIC framework to measurable constructs. The items were developed through conceptual grounding in prior TIC and related HCI/healthcare literature, followed by iterative refinement based on expert feedback to improve clarity and relevance to chatbot experiences [27, 31, 39–41].

The questionnaire was developed as a self-designed instrument and refined with inputs from an expert in clinical psychology with expertise in TIC (the health system the authors are from), as well as feedback from human-centered design experts with experience in mental health contexts and AI ethics. These inputs helped improve the clarity and relevance of the items while ensuring alignment with trauma-informed principles. Likert-scale response formats were used to allow participants to indicate the degree of their experiences or agreement in a consistent and interpretable manner across constructs. Item wording was adapted to reflect user-facing experiences with chatbots rather than clinical or provider-oriented language. [26, 27, 34, 42].

The survey comprised four primary content sections:

1. **Demographics**, which captured participants' gender, age, race/ethnicity, education level, and history of adverse life experiences;
2. **Technology Experience**, which assessed participants' proficiency with smartphones, artificial intelligence, and smart devices, as well as their motivations for using mental health chatbots;
3. **Trauma-Informed Chatbot Characteristics and Experiences**, which evaluated user perceptions of TIC principles as reflected in chatbot design and interaction quality; and
4. **Satisfaction Level**, which measured overall satisfaction and perceived trauma-informed quality of the chatbot interaction.

The survey in total includes 59 items, including multiple-response (checkbox), Likert-scale (rated from 1 = not good/proficient/never to 5 = very good/proficient/always), binary (yes/no), and open-ended questions. To ensure data quality, the survey included one attention-check item and two hidden bot-detection questions designed to identify inattentive or automated responses. While participation in all questions was not mandatory and respondents could skip items at their discretion, they were required to provide responses to demographic items and multiple-choice questions related to TIC. Participants received $10 in compensation for survey completion. Full survey items are provided in Supplementary Table S1.

2.2 Recruitment

Participants were recruited between December 2024 and April 2025 through three primary channels: the crowdsourcing platform Clickworker ($n = 488$), online social media platforms ($n = 113$), and the MyChart patient communication portal ($n = 5$), yielding a final eligible sample of 606 individuals [43]. Participation was voluntary, and each participant received $10 as compensation for their time.

Clickworker, a trusted crowdsourcing platform, was selected as the primary recruitment channel given its access to a large, diverse, and verified workforce spanning a broad range of educational backgrounds [43]. Crowdsourcing platforms, such as Clickworker, are widely used in human-computer interaction and digital health research to recruit diverse, scalable participant samples [44]. To address common data-quality concerns associated with crowdsourced samples, several safeguards were implemented, including explicit attention-check items (e.g., "To make sure you are paying attention, please select the third option for this question"), embedded bot-detection questions (e.g., "What is your favorite color?" and "What is your favorite ice cream flavor"), and post hoc exclusion of respondents who reported using tools that did not meet the definition of a mental health chatbot. Only participants who successfully passed all quality checks were included in the final analytic sample.

Social media recruitment was conducted via Twitter and LinkedIn, where study information and a survey link were publicly posted. The same data-quality safeguards applied to the crowdsourced sample were implemented for social media respondents.

Recruitment via MyChart targeted Indiana-based patients aged 18 years or older who had at least one clinical encounter within the behavioral health or psychiatry departments of the authors' health system in the preceding 12 months and an active behavioral or psychiatric episode documented in their record. Although recruitment for MyChart spanned approximately eight weeks, the resulting subsample ($n = 5$ participants) was small and is therefore not interpreted as representative of a clinical population, likely reflecting the sensitivity of trauma-related research topics and the well-documented reluctance of clinical populations to participate in such studies

The study was not designed to exclusively recruit trauma-affected or help-seeking clinical populations; instead, recruitment was intentionally broad to reflect the real-world diversity of individuals, both with and without trauma exposure, who engage with digital mental health tools. This approach aligns with an increasingly recognized perspective in human-computer interaction and mental health technology research

that such heterogeneous user populations are a valid and important focus of study. Accordingly, the study is framed as exploratory and user-centered rather than intended for clinical generalization.

2.3 Participants

A total of 606 individuals constituted the final eligible analytic sample. Eligibility required participants to be: (1) aged 18 or older; (2) self-reporting a history of mental health concerns; (3) having used at least one mental health chatbot; and (4) able to read and understand English. The informed consent item was displayed only to individuals satisfying all four criteria. Those who failed eligibility screening, quality checks, or provided invalid chatbot responses were excluded.

Participants were instructed to reflect on their most recent or most frequently used mental health chatbot, acknowledging that perceptions may reflect aggregated experiences across heterogeneous platforms. This approach aligns with the study's objective of assessing trauma-informed characteristics of mental health chatbots at a conceptual level, while acknowledging that platform-level variability may influence factor structure and outcome estimates.

Overall, 62.4% of participants were female, 37.5% were aged 25–34 years, 75.2% identified as White, and 30.2% held a bachelor's degree. A majority reported a history of household dysfunction (71.3%), followed by experiences of various forms of abuse. Detailed demographic characteristics of the final analytic sample are presented in Table 1.

2.4 Data analysis

Qualitative data from open-ended survey questions were analyzed using inductive reflexive thematic analysis to identify recurring patterns and user-defined concepts related to trauma-informed chatbot experiences. Reflexivity was maintained throughout by the first and last authors to ensure transparency in interpretive decision-making [45].

Quantitative survey data, including multiple-response and Likert-scale items, were analyzed using a combination of descriptive and inferential statistical methods. Descriptive statistics were first computed to summarize participant characteristics, technology use patterns, and overall perceptions of trauma-informed care. Exploratory factor analysis (EFA) was then conducted to identify latent domains underlying participants' conceptualization of trauma-informed chatbot characteristics. Confirmatory factor analysis (CFA) was subsequently performed to validate the factor structure and evaluate model fit. Finally, multivariable logistic regression models were fitted to identify factors associated with participants' perceptions of chatbot interactions as trauma-informed.

All statistical analyses were conducted using R (version 4.4.3) and relevant statistical packages. An overview of the analytic workflow used to address the study research questions is presented in Figure 1.

2.4.1 Descriptive analysis

Descriptive statistics were computed to summarize participants' overall satisfaction, technology proficiency, chatbot use and motivations, as well as multiple-response items on TIC perceptions. Frequency distributions and proportions were reported for categorical and binary variables, while means and standard deviations were calculated for continuous and ordinal variables.

2.4.2 Factor analysis

To identify the underlying structure of TIC perceptions, an EFA was conducted on the full set of Likert-scale items from the TIC section of the survey. EFA was chosen given the exploratory aim of identifying latent constructs underlying participant responses, as no established factor structure was assumed. Prior to analysis, all variables were assessed for missing data. The average item-level missing rate was $< 2\%$. To preserve the integrity of the original responses, no imputation was performed; instead, cases with missing data were excluded listwise from the analysis. EFA was conducted using maximum likelihood estimation on the correlation matrix, with an oblimin rotation applied to account for potential correlations among factors. The number of factors retained was guided by parallel analysis. A stepwise item reduction procedure was applied, sequentially removing the item with the lowest standardized loading at each iteration until all retained items demonstrated primary factor loadings ≥ 0.50 to ensure strong associations with the extracted factors. Following item selection, the theoretical framework was refereed to ensure retained items remained conceptually coherent. CFA was subsequently conducted to test the fit of the TIC model identified in the EFA phase. CFA models were evaluated using standard goodness-of-fit indices, including the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR).

Because EFA and CFA were performed on the same sample, results are exploratory and descriptive, intended to examine internal structure rather than definitive model validation. Nevertheless, the combination of theory-informed item development, stringent loading thresholds, oblique rotation, and multiple fit criteria supports the internal coherence and preliminary construct validity of the proposed TIC factor structure. Future research should replicate and validate this model in independent, clinically representative samples.

2.4.3 Logistic regression model

Logistic regression analyses were conducted to examine predictors of whether participants perceived their chatbot experience as trauma-informed. Composite scores were first calculated from the validated CFA structure, with items loaded on the same latent factor averaged to form summary scores for each dimension. These composite scores were merged with demographic variables and technology proficiency data to serve as predictors in a multivariable logistic regression model. A forward stepwise selection procedure was used to identify the most parsimonious model. Final model estimates are reported as odds ratios (ORs) with corresponding p-values and 95% confidence

10.3389/fdgth.2026.1797681

TABLE 1 Demographic characteristics of participants by recruitment source (*N* = 606).

Characteristics	Clickworker (*N* = 488)	Social media (*N* = 113)	MyChart (*N* = 5)	Overall (*N* = 606)
Gender				
Female	326 (66.8%)	47 (41.6%)	5 (100.0%)	378 (62.4%)
Male	152 (31.1%)	65 (57.5%)	0 (0.0%)	217 (35.8%)
Nonbinary	8 (1.6%)	0 (0.0%)	0 (0.0%)	8 (1.3%)
Transgender	1 (0.2%)	1 (0.9%)	0 (0.0%)	2 (0.3%)
Prefer not to say	1 (0.2%)	0 (0.0%)	0 (0.0%)	1 (0.2%)
Age group				
18–24 years	37 (7.6%)	12 (10.6%)	0 (0.0%)	49 (8.1%)
25–34 years	174 (35.7%)	51 (45.1%)	2 (40.0%)	227 (37.5%)
35–44 years	183 (37.5%)	41 (36.3%)	1 (20.0%)	225 (37.1%)
45–54 years	87 (17.8%)	9 (8.0%)	2 (40.0%)	98 (16.2%)
55–64 years	7 (1.4%)	0 (0.0%)	0 (0.0%)	7 (1.2%)
65 years or older	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Race				
White	373 (76.4%)	79 (69.9%)	4 (80.0%)	456 (75.2%)
Black	77 (15.8%)	23 (20.4%)	0 (0.0%)	100 (16.5%)
Asian	37 (7.6%)	3 (2.7%)	0 (0.0%)	40 (6.6%)
American Indian	15 (3.1%)	7 (6.2%)	0 (0.0%)	22 (3.6%)
Pacific Islander	2 (0.4%)	2 (1.8%)	0 (0.0%)	4 (0.7%)
Other	18 (3.7%)	0 (0.0%)	1 (20.0%)	19 (3.1%)
Missing	1 (0.2%)	0 (0.0%)	0 (0.0%)	1 (0.2%)
Adverse life experiences				
Abuse	323 (66.2%)	53 (46.9%)	5 (100.0%)	381 (62.9%)
Adoption	19 (3.9%)	6 (5.3%)	1 (20.0%)	26 (4.3%)
Child services involvement	79 (16.2%)	7 (6.2%)	2 (40.0%)	88 (14.5%)
Community adversity	198 (40.6%)	39 (34.5%)	3 (60.0%)	240 (39.6%)
Household dysfunction	372 (76.2%)	55 (48.7%)	5 (100.0%)	432 (71.3%)
Neglect	234 (48.0%)	50 (44.2%)	4 (80.0%)	288 (47.5%)
Education level				
Less than high school	12 (2.5%)	0 (0.0%)	0 (0.0%)	12 (2.0%)
High school or GED	94 (19.3%)	3 (2.7%)	0 (0.0%)	97 (16.0%)
Some college (no degree)	129 (26.4%)	5 (4.4%)	2 (40.0%)	136 (22.4%)
Technical certification	25 (5.1%)	3 (2.7%)	0 (0.0%)	28 (4.6%)
Associate degree	55 (11.3%)	26 (23.0%)	2 (40.0%)	83 (13.7%)
Bachelor's degree	127 (26.0%)	55 (48.7%)	1 (20.0%)	183 (30.2%)
Master's degree	41 (8.4%)	14 (12.4%)	0 (0.0%)	55 (9.1%)
Doctoral degree	2 (0.4%)	7 (6.2%)	0 (0.0%)	9 (1.5%)
Prefer not to say	2 (0.4%)	0 (0.0%)	0 (0.0%)	2 (0.3%)
Missing	1 (0.2%)	0 (0.0%)	0 (0.0%)	1 (0.2%)

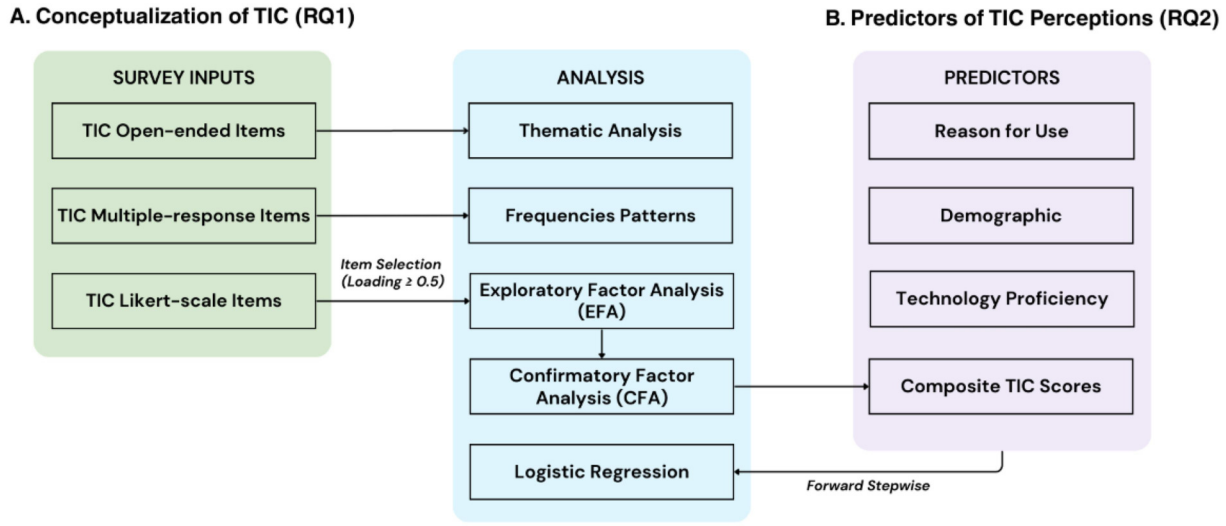


FIGURE 1
Overview of the analytic workflow used to address the study research questions.

intervals (CIs), providing insight into which chatbot features are most strongly associated with positive TIC perceptions.

3 Result

3.1 General descriptive statistics

3.1.1 Overall satisfaction

A total of 606 surveys were included in analysis. Overall, 92.9% of respondents perceived the chatbot as trauma-informed, with similarly high endorsement across recruitment sources (Clickworker: 91.6%, social media: 100.0%, MyChart: 60.0%). Likewise, 92.6% of participants reported they would recommend the chatbot to others with a trauma history (Clickworker: 91.0%, social media: 100.0%, MyChart: 80.0%). Satisfaction ratings further supported chatbot acceptability. On a five-point scale ranging from "very dissatisfied" to "very satisfied," most participants expressed a positive experience, with 84.5% selecting the top two satisfaction levels (48.0% rated it as satisfied and 36.5% as extremely satisfied). Lower satisfaction ratings were rare, with only 2.9% of respondents expressing dissatisfaction. These patterns were generally consistent across all recruitment sources.

3.1.2 Technology experience

Participants demonstrated generally high levels of self-reported proficiency across all six technology domains: smartphone use, internet browsing, AI tools, computer use, social media, and smart home devices. Among these, smartphone use (M = 4.73, SD = 0.54) and internet browsing (M = 4.71, SD = 0.56) proficiencies were rated highest across the full sample. In contrast, proficiency in smart home device use was the lowest overall (M = 3.99, SD = 1.02), reflecting broader variability in exposure to and use of emerging technologies. Clickworker respondents reported the highest average proficiency overall (M = 4.44, SD = 0.82), followed by social media participants (M = 4.37, SD = 0.86). MyChart participants (M = 4.10, SD = 0.71) had slightly lower average scores across most domains. Despite these differences, mean scores remained high, indicating strong baseline digital literacy among participants. The details can be seen in Supplementary Figure S1.

3.1.3 Chatbot use and motivations

Participants reported varied engagement with mental health chatbots in terms of both frequency and platform type. The results are presented in Figure 2. Most used the chatbot 1–3 times per month (19.3%) or 2–5 times per week (19.0%), while 17.5% reported trying it only once. Daily use was less common (8.4%), indicating that chatbots may be serving as intermittent support tools rather than daily companions for most users. For platforms, the most frequently used chatbots were Serenity (23.1%), Woebot (21.5%), and Stresscoach (19.8%). Participants reported diverse motivations for engaging with mental health chatbots. The most endorsed reasons were related to emotional distress, including stress or anxiety (68.8%) and depression (59.4%). A substantial proportion also reported exploratory motives such as self-guided exploration (46.0%) and curiosity (37.8%). The desire for a psychologically safe space was also a prevalent motivation (41.1%), ranking higher than expected clinical drivers such as trauma (32.5%) or loneliness (36.3%). Less commonly selected reasons included privacy concerns, recommendations, and previous experience, each under 20%.

3.2 RQ1: conceptualization of TIC principles

3.2.1 Users perceptions across six core dimensions

To explore how users conceptualized trauma-informed care (TIC) in the context of mental health chatbots, participants'

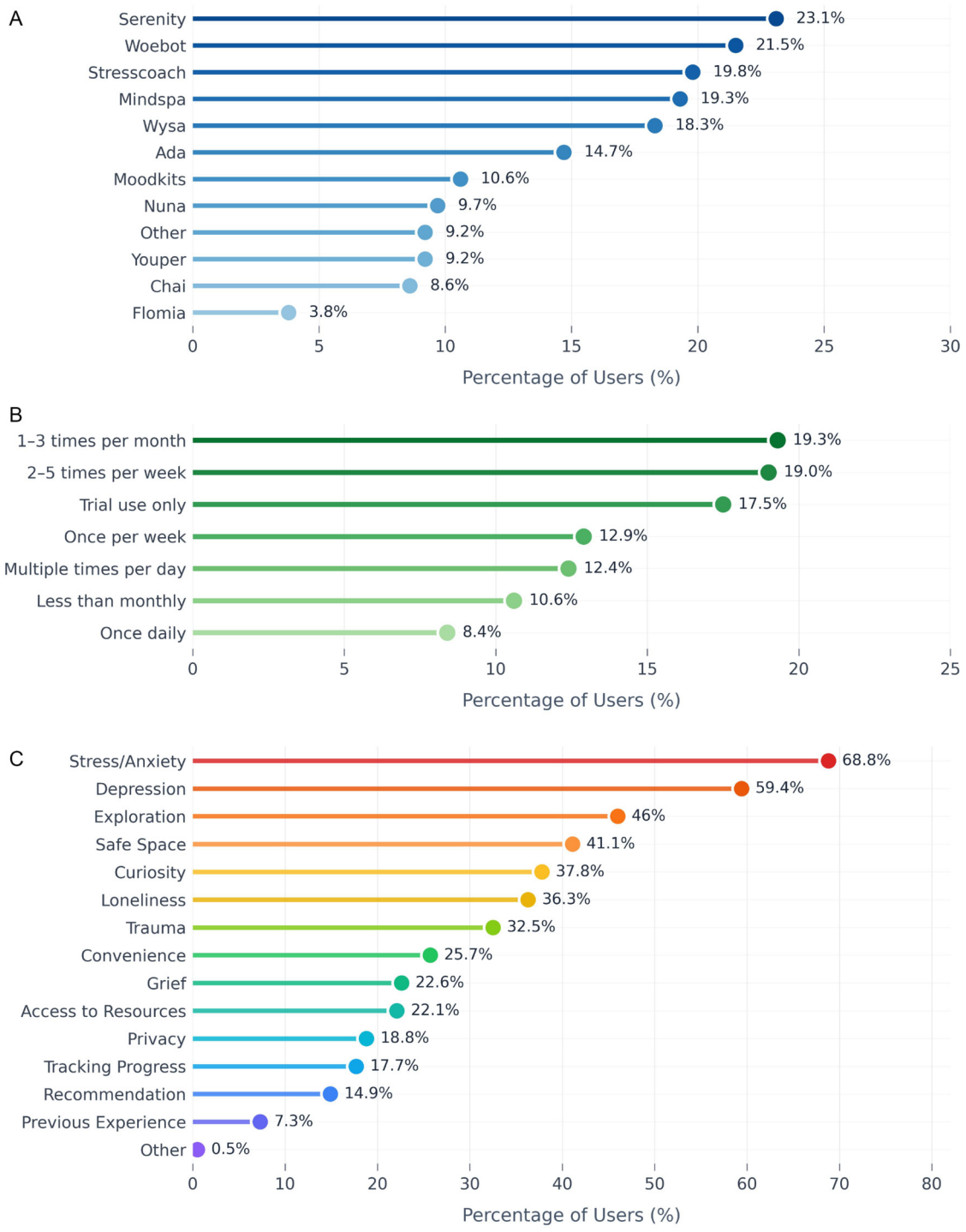


FIGURE 2
Patterns of mental health chatbot use among participants. **(A)** Distribution of mental health applications used. **(B)** Self-reported frequency of application usage. **(C)** Primary reasons for using mental health applications. Respondents could select multiple options; percentages reflect the proportion of the total sample.

responses to multiple-response checkbox items aligned with the six core TIC principles were summarized. Both individual feature frequencies and co-occurrence patterns were examined to characterize how users collectively defined a trauma-informed chatbot.

Safety was primarily defined as the ability to interact with the chatbot without fear or judgment. Emotional safety was the most frequently selected feature within this domain (86.7%), highlighting the importance of psychological security in chatbot interactions. The most common feature combination (28.9%) included emotional safety, physical safety, and digital safety, indicating a holistic view of safety spanning psychological and technical dimensions.

Trust and transparency centered on ensuring that users feel respected through open and clear communication. Information

confidentiality emerged as the highest-priority feature (82.3%), underscoring strong concerns regarding privacy. A frequently occurring combination (26.7%) included transparency, tailored responses, validation of experiences, and data care, suggesting that trust is closely linked to both personalization and responsible data handling.

Empowerment emphasized user autonomy and support. Guide conversation was the most frequently endorsed feature (73.1%), underscoring the importance of user-directed dialogue. The most common combination (20.8%) included sharing control, no forced choice, expression options, hopeful messaging, and sense of agency, reflecting preferences for tools that foster autonomy and emotional encouragement.

Collaboration reflected partnership and shared decision-making. Respect choices was the most frequently selected design element (72.6%). A commonly endorsed combination (38.8%) included goal setting, shared goals, and respect choices, suggesting that participants valued mutuality and co-creation of a therapeutic path.

Peer support focused on connection with others who have similar lived experiences. Ask for support was the most frequently selected feature (73.1%), highlighting the importance of self-initiated outreach. The most common configuration (19.8%) included offer support, support available, relevant connections, and virtual or in-person support, indicating interest in accessible and meaningful peer engagement.

Cultural sensitivity emphasized inclusivity and respect for diverse identities. Nonjudgmental was the most frequently endorsed feature (79.4%), reflecting a strong desire for acceptance. The most common feature combination (23.8%) included sensitive to identity, avoids offense, avoids triggers, validates experience, shows interest, and culturally aligned, highlighting preferences for chatbots that are affirming, inclusive, and mindful of cultural and individual differences.

Full item-level response distributions are presented in Supplementary Figure S2, and detailed item wording is provided in Supplementary Table S2.

3.2.2 Uncovering latent constructs of trauma-informed experience

Participants' conceptualization of trauma-informed chatbot design was further explored through Likert-scale ratings, in which they indicated their level of agreement (strongly disagree to strongly agree) with statements reflecting the six TIC principles. This measure captured the extent to which users endorsed statements such as, "My chatbot is sensitive to my concerns, identity, and/or historical experiences." Overall, responses reflected favorable perceptions: 75.2% of participants selected a positive response (Agree: 35.5%; Strongly Agree: 39.7%), 16.7% were neutral, and only a small minority expressed disagreement (Disagree: 5.6%; Strongly Disagree: 2.6%). Item-level response distributions are presented in Supplementary Figure S3.

While these descriptive findings suggest broadly positive user perceptions, prior HCI research cautions that high endorsement alone does not reveal how users cognitively organize or differentiate experiential constructs in human–AI interaction [46]. Accordingly, we conducted an EFA to examine whether participants' responses aligned with the six theoretically specified TIC principles or reflected an alternative latent structure shaped by the interactional affordances of chatbot systems.

Parallel analysis supported a five-factor solution. The final EFA model retained 18 items and explained 66% of the total variance, with individual factors accounting for 7% to 20% of the explained variance. Based on item content and theoretical coherence, the five factors were labeled *Peer Support*, *Trust and Transparency*, *Data Safety*, *Empowerment*, and *Cultural Sensitivity*. These factors represent the most salient trauma-informed dimensions as experienced by users during real-world chatbot interactions.

Four of the five factors mapped closely to SAMHSA TIC principles. *Peer Support* reflected users' desire for connection to others with shared lived experiences, consistent with prior findings that perceived social connectedness, whether direct or mediated, is central to supportive digital mental health experiences [47]. *Trust and Transparency* encompassed perceptions of reliability, expectation alignment, and responsiveness, echoing HCI literature demonstrating that trust in conversational agents is shaped by predictability and clarity rather than anthropomorphism alone [48]. *Empowerment* captured feelings of agency, autonomy, and emotional self-efficacy, aligning with self-determination theory and prior work on empowering AI design [49, 50]. *Cultural Sensitivity* reflected inclusive language use and avoidance of triggering content, consistent with research highlighting linguistic safety as a key determinant of perceived care quality in digital mental health tools [51]. *Data safety* emerged as a distinct latent construct, encompassing items related to digital safety, privacy being respected, and careful handling of personal data, rather than as part of a broad, undifferentiated safety factor.

The SAMHSA principle of *Collaboration* did not emerge as a distinct factor. Instead, collaboration-related items loaded onto adjacent domains, particularly *Empowerment* and *Trust and Transparency*. This pattern suggests that users may conceptualize collaboration with chatbots less as reciprocal partnership and more as feeling supported, respected, and enabled. Factor correlations ranged from 0.31 to 0.68, indicating related but distinct constructs, and mean item complexity was low (1.1), supporting factor interpretability. Model fit was acceptable (RMSR = 0.02; $\chi^2 = 146.61$, $df = 73$, $p < .001$), and internal consistency was strong across factors (Cronbach's $\alpha = 0.77$–0.92). Detailed factor loadings and communalities are reported in Supplementary Table S3.

3.2.3 Validating the factor structure of trauma-informed experience

To evaluate the adequacy of the factor structure identified through EFA, a CFA was conducted using maximum likelihood estimation. The five-factor, 18-item model demonstrated excellent fit to the data (CFI = 0.978; TLI = 0.973; RMSEA = 0.045, 90% CI: 0.037–0.052; SRMR = 0.033). All standardized factor loadings were statistically significant ($p < .001$) and ranged from 0.70 to 0.88, indicating strong associations between observed indicators and their respective latent constructs.

Among the five latent domains confirmed through CFA, Peer Support emerged as a particularly cohesive and robust construct,

with nearly all items demonstrating strong factor loadings. Most notably, the item "Q42. Support Available: My mental health chatbot offers opportunities for peer support and community connections as part of my mental health management" exhibited the highest standardized loading in the entire model (0.88), establishing it as a central indicator of this domain.

Within the Trust and Transparency domain, the item "Q22. Meets Expectations: My chatbot responds to my experiences in a way that meets my expectations" had the highest loading (0.81), highlighting its critical role in shaping user perceptions of trustworthiness and responsiveness. In the Data Safety domain, "Q15. Privacy Respected: My mental health chatbot respects my privacy and confidentiality during interactions" demonstrated the strongest loading (0.83), affirming that users prioritize digital privacy and data protection as foundational to feeling safe in chatbot interactions. For Empowerment, the leading item was "Q34. Sense of Agency: My chatbot helps me believe that I have agency/ability to change things in my life," which yielded a loading of 0.88, underscoring the importance of fostering personal autonomy and emotional self-efficacy. Lastly, within Cultural Sensitivity, "Q51. Avoids Triggers: My chatbot avoids words or phrases that may trigger me" showed the strongest association (0.81), emphasizing users' sensitivity to emotionally safe language.

Figure 3 presents a synthesized summary of findings, integrating both the descriptive checkbox responses and the CFA results. The detailed standardized loadings for each item within the CFA model are presented in Figure 4 and Supplementary Table S4, supporting the scale's construct validity and further clarifying item-factor relationships.

3.2.4 Qualitative evidence supporting the five-factor structure

Qualitative findings further supported the five-factor structure identified through EFA and CFA by illustrating how users experienced each TIC domain in practice. Across domains, participants' narratives provided concrete examples of how abstract TIC principles were enacted during chatbot interactions, thereby reinforcing the quantitative findings.

Within the **Trust and Transparency** domain, participants consistently emphasized the importance of emotional validation and responsive communication. Users described feeling understood by the chatbot, noting that it would "listen and understand," and that "I trust it more because it feels like I'm talking with a caring and concerned friend, not a chatbot that just gives basic random responses." These accounts highlight how perceived empathy and tailored responses foster trustworthiness in automated mental health support.

Data Safety was described as a holistic construct encompassing anonymity, immediacy of access, and digital protections. Several participants identified anonymity as central to their sense of safety, with one noting that "the main thing that helps me feel safe" was the ability to interact without revealing personal identity. Others emphasized the importance of data protection, explaining that "knowing my conversations are secure and private makes me feel safer." Together, these responses illustrate how technical safeguards contribute to users' perceptions of safety.

Experiences related to **Empowerment** centered on agency, encouragement, and goal support. Participants expressed appreciation for features that promoted self-direction and motivation, with one reporting that "consistent check-ins help keep me on track with my goals." These reflections align with the quantitative findings, underscoring the role of empowerment in shaping trauma-informed perceptions.

Responses regarding **Peer Support** were more heterogeneous. While some participants preferred private, one-on-one chatbot interactions, others expressed interest in shared experiences, noting that "peer-to-peer interactions... would be appreciated sometimes." This variability suggests that peer support is valued selectively and may benefit from flexible, opt-in design approaches.

Finally, **Cultural Sensitivity** emerged as a critical dimension of inclusion and respect. Participants highlighted the importance of consent, nonjudgmental language, and attentiveness to emotional boundaries, stating that "the bot does check in to see

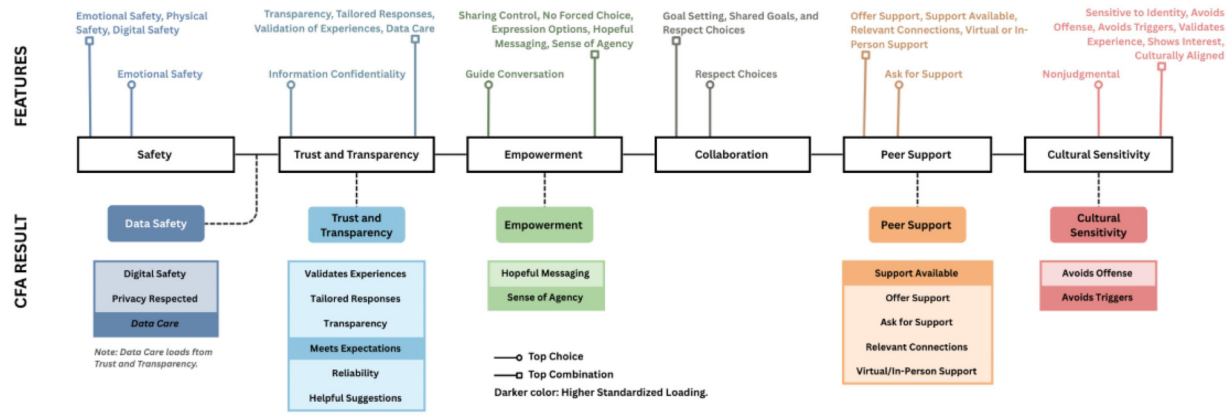


FIGURE 3
Summary of descriptive and confirmatory factor findings on trauma-informed chatbot features. The top portion summarizes participants' most frequently endorsed features (circle endpoints) and commonly co-selected combinations (square endpoints) across the original six trauma-informed care principles. The bottom portion displays the final five-factor structure derived from confirmatory factor analysis, with darker colors indicating items with higher standardized factor loadings.

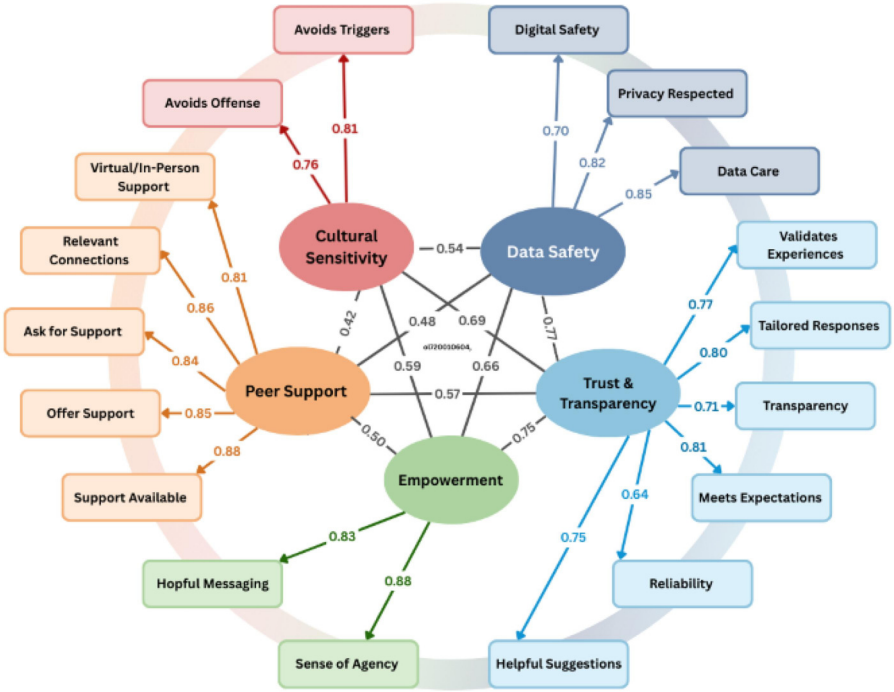


FIGURE 4
Factor structure with standardized loadings for the Confirmatory Factor Analysis model of trauma-informed chatbot perception.**TrS**: Trust; **Sft**: Safety; **Emp**: Empowerment;**PrS**: Peer Support; **SnS**: Cultural Sensitivity. Bar widths represent the magnitude of standardized loadings, with values above 0.80 indicating particularly strong associations between items and their respective latent constructs.

if I am comfortable" and "asks permission to discuss a sensitive topic further." These accounts emphasize users' expectations for culturally and emotionally attuned communication.

3.3 RQ2: identifying predictors for chatbot being trauma-informed

To examine factors associated with perceiving a mental health chatbot as trauma-informed, we fitted a multivariable logistic regression model. Composite scores for the five CFA-derived dimensions were first entered into a baseline model. Forward stepwise selection was subsequently applied to identify a parsimonious set of predictors. Candidate variable *reason: others* were excluded from the final model due to estimation instability (e.g., near-infinite standard errors), driven by sparse category endorsement and small subgroup sizes. The final model results are summarized in Figure 5.

3.3.1 Chatbot design features

Among design-related predictors, perceived *Trust* emerged as the strongest and most stable correlate of trauma-informed perception. Higher trust scores were associated with substantially increased odds of perceiving the chatbot as trauma-informed (OR = 3.89, 95% CI: 1.73–8.77, p = .001). *Peer Support* (OR = 1.73, 95% CI: 1.09–2.74, p = .021) and *Empowerment* (OR = 1.97, 95% CI: 1.09–3.57, p = .025) were also positively associated with trauma-informed perception, reinforcing the importance of relational responsiveness and autonomy-supportive features in shaping users' evaluations of chatbot interactions.

3.3.2 User characteristics

Several user-level characteristics further differentiated trauma-informed perceptions. Participants who reported *prior involvement with the Department of Child Services* exhibited markedly higher odds of perceiving the chatbot as trauma-informed (OR = 9.71, 95% CI: 1.66–56.79, p = .012). Although confidence intervals were wide, reflecting limited subgroup size, this finding suggests that individuals with lived experience of formal care systems may be particularly attuned to trauma-aligned design features.

In contrast, two predictors were negatively associated with trauma-informed perception. Users who cited *convenience* as their primary motivation for chatbot use had significantly lower odds of perceiving the chatbot as trauma-informed (OR = 0.33, 95% CI: 0.11–0.96, p = .041). Similarly, higher self-rated *smartphone proficiency* was inversely associated with trauma-informed perception (OR = 0.38, 95% CI: 0.16–0.93, p = .033). These associations may reflect differing evaluative standards among highly digitally proficient users, who may prioritize efficiency, performance, and usability over relational or emotionally supportive features, and therefore be less responsive to trauma-informed design cues.

4 Discussion

Our study offers a rigorous empirical foundation for understanding how users interpret, experience, and evaluate TIC features in mental health chatbot interactions. By integrating descriptive pattern analysis, confirmatory factor analysis, and

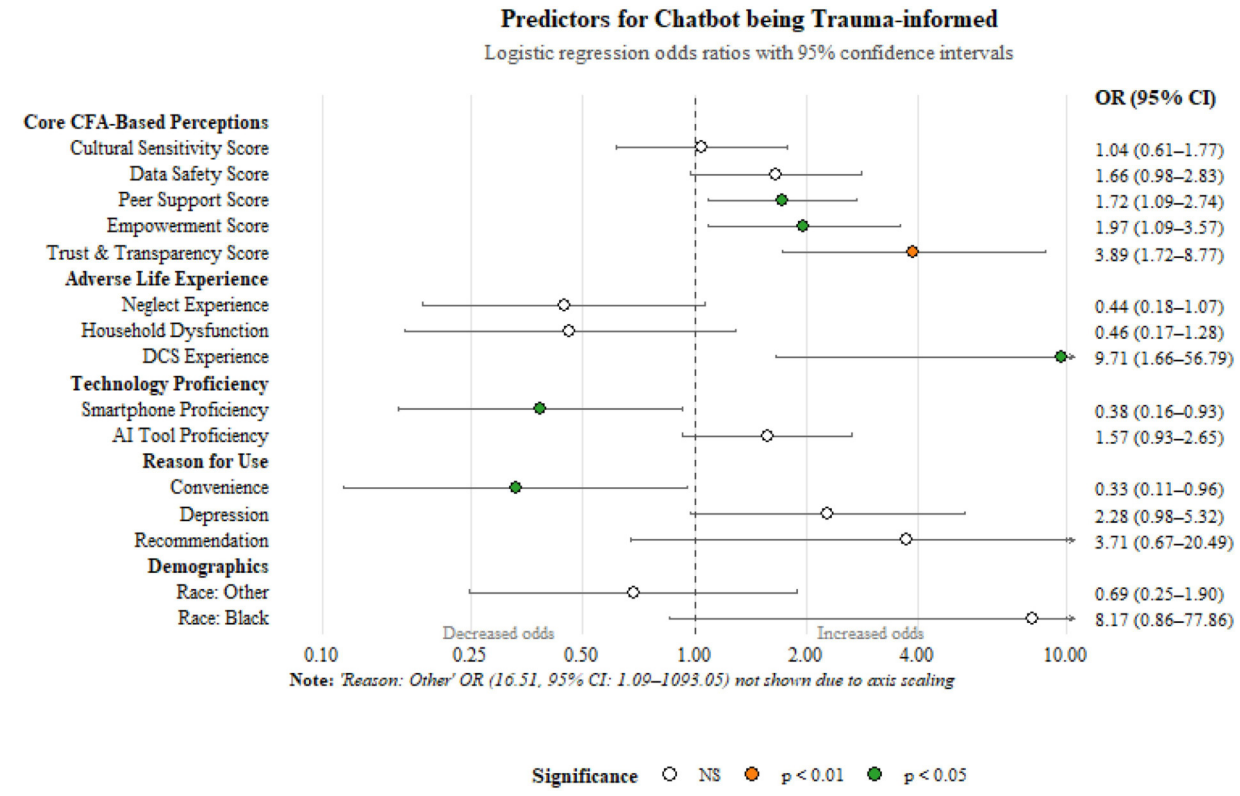


FIGURE 5
Multivariable logistic regression results showing predictors of perceived trauma-informed chatbot interaction. Estimates are presented as odds ratios with 90% confidence intervals and corresponding *p*-values.

predictive modeling, we elucidate both the conceptual structure of TIC as perceived by end-users and the key elements that drive trauma-informed appraisals.

Overall, users overwhelmingly perceived the chatbot as embodying TIC characteristics, with more than 90% endorsing their experience as trauma-informed alongside similarly high rates of satisfaction and recommendation. Importantly, participants did not interpret TIC as a monolithic construct. Instead, both quantitative and qualitative findings revealed a nuanced understanding spanning emotional safety, privacy, identity validation, agency, and peer connection. These themes align with existing TIC theory while contributing new empirical evidence to its application in digital mental health contexts [26, 27, 34, 42]. Rather than viewing TIC as a single construct, participants emphasized key design features that shaped how they judged the chatbot's empathy, safety, and trustworthiness. Emotional Safety and Information Confidentiality were most valued, confirming that freedom from fear, judgment, and privacy concerns defines psychological safety in digital care [2, 19, 20, 52]. Features that allowed users to guide conversations and receive hopeful, personalized responses fostered empowerment and a sense of agency [52, 53]. Respect for user choices and adaptive feedback conveyed collaboration and mutuality without explicit co-decision-making [29, 54, 55]. Opportunities for self-initiated peer connection further enhanced perceived support and belonging [47, 56]. Finally, a nonjudgmental tone and cultural awareness reinforced inclusivity and emotional comfort across backgrounds [39]. Together, these design elements illustrate that users interpret trauma-informed chatbots through an integrated lens of emotional security, autonomy, relational responsiveness, and cultural respect.

Our factor analysis results yielded a five-factor structure comprising Trust and Transparency, Data Safety, Empowerment, Peer Support, and Cultural Sensitivity. The factor analysis revealed a distinct Data Safety latent factor, which suggests that, for our population, safety was primarily understood and evaluated through the lens of data protection and privacy, rather than a generalized sense of emotional or interpersonal safety. In the context of mental health chatbots, this distinction indicates that assurances about how user data are stored, protected, and used may be especially salient for individuals with lived experiences of vulnerability or trauma and may meaningfully shape their perceptions of whether an interaction feels safe. The absence of a distinct Collaboration factor, despite its theoretical salience, suggests that collaborative elements (e.g., shared decision-making) may be implicitly embedded within other domains. Qualitative accounts reinforced this interpretation: several participants emphasized valuing "consistent feedback" and the sense that responses were "tailored to me," underscoring collaboration as an integrated rather than standalone dimension.

Trust and Transparency emerged as the most robust driver of trauma-informed perception, especially in relation to confidentiality and expectation management. Emotional validation played a central role in cultivating trust, as reflected

in both model estimates and user narratives. Participants consistently described the importance of feeling "listened to and understood" and noted that empathetic language made them feel "like I'm talking with a caring and concerned friend, not just a chatbot." Such validation highlights how relational design elements underpin trust and drive trauma-informed evaluations. This finding aligns with prior research demonstrating that relational authenticity and transparent communication are essential for trust-building in digital mental health tools [19, 20].

Empowerment and Peer Support also proved to be critical dimensions for shaping trauma-informed perceptions. Users who felt supported in decision-making or had access to community-based resources were significantly more likely to view the chatbot as trauma-informed, echoing findings by Potts et al., who observed that multilingual chatbots offering user-driven dialogue improved well-being and engagement [53]. Similarly, participants valued access to supportive, community-oriented features, paralleling Beilharz et al., who emphasized that peer-based and family-linked chatbot components enhanced feelings of connection and emotional safety among adolescents [56]. This aligns with core TIC principles that emphasize relational healing and shared agency as foundational to care [4, 27].

Individual differences further shaped TIC perceptions. Participants in our study with a history of involvement with child protective or social services had significantly higher odds of perceiving the chatbot as trauma-informed.This finding may be understood through the lens of system-impacted populations' unique relationships with institutional support and technology. Research on youth with foster care or child welfare involvement has documented that such experiences often lead to distinctive patterns of help-seeking behavior and trust formation [57–60]. Youth in foster care frequently experience multiple placements and disrupted relationships, which can profoundly shape their expectations and evaluations of new support systems, whether human or technological [58, 60–63]. Studies examining digital health adoption among foster youth suggest that this population approaches technology-mediated support with both strategic caution and openness, particularly valuing features that signal safety, autonomy, and non-coercive care [59, 60, 64]. These individuals often develop what might be termed "system literacy": a heightened ability to discern whether systems, including digital ones, are designed with their safety and dignity in mind [61–63]. Prior work has also shown that young people who have experienced foster care, adoption, or removal from biological families may be particularly attuned to relational qualities in both human and technological interactions, given their histories of disrupted attachments and mandated institutional relationships [58, 65–69].

Consistent with this interpretation, research on relational agents has shown that users experiencing higher emotional vulnerability, including depressive symptoms or psychological distress, tend to form stronger therapeutic alliances with conversational systems [70, 71] For individuals with child welfare histories, this effect may be compounded: digital agents may represent a form of support perceived as less surveilling, less coercive, and more user-controlled than traditional institutional services [57, 61, 63, 72]. These studies suggest that conversational agents may be experienced as emotionally safer or less threatening than human interlocutors, in part because they offer predictable, nonjudgmental interactions and reduce fears of stigma or social evaluation [61, 62, 72]. Youth in foster care often report mistrust of service providers and concerns about stigma when seeking mental health support , barriers that may be attenuated in interactions with digital agents [61, 63, 72]. Similar findings have been reported in digital mental health contexts, where users with greater distress often report higher perceived empathy and trust in chatbot-based support tools [30, 73]. Moreover, research specifically examining technology use among foster care populations indicates that digital platforms may offer a sense of privacy and agency that is particularly valued by those who have experienced institutional oversight and limited control over their lives [57, 60, 74].

Conversely, users who cited convenience as their primary reason for chatbot use, or who rated themselves highly proficient with smartphones, were significantly less likely to perceive the chatbot as trauma-informed. This pattern aligns with prior work suggesting that technologically experienced or efficiency-oriented users tend to evaluate conversational systems through a utilitarian lens, prioritizing speed, task completion, and functional accuracy over relational or emotional qualities [75, 76]. Additionally, higher smartphone proficiency may foster greater critical awareness of automated empathy, leading some users to approach relational or caring language with increased skepticism [77]. Related work in digital self-care and health technologies has similarly shown that users motivated by convenience often place lower value on emotional depth, which can lead to reduced perceptions of empathy or trust [54]. In such cases, trauma-informed features, such as validation, reflective language, or user control, may be perceived as peripheral or insufficient relative to expectations for seamless performance.

Overall, our results offer critical insights for the future of trauma-informed AI in mental health. First, we demonstrate that TIC is both legible and empirically measurable: users overwhelmingly endorsed trauma-informed experiences, and factor analysis revealed a robust five-domain structure that provides strong evidence for measurement accuracy. Second, predictive modeling identified high-impact predictors of trauma-informed perceptions, spanning both chatbot design features (e.g., trust, empowerment, and peer support) and user characteristics (e.g., trauma history, motivations for use, and technology proficiency), thereby underscoring the importance of personalization and user-centered design. Third, our findings reveal meaningful variation in how different user populations perceive and prioritize TIC features, suggesting that one-size-fits-all approaches may inadequately serve diverse trauma survivors. Designers of mental health chatbots should embed trauma-informed principles not as optional enhancements but as core architectural features. Likewise, TIC-aligned metrics should be systematically incorporated into evaluation frameworks. As reliance on automated tools for mental health support continues to grow, establishing rigorous standards for safety, ethics, and human-centered care will be essential.

Several limitations should be considered when interpreting these findings. First, the study employed a cross-sectional, self-reported survey design, which may be influenced by recall bias or social desirability effects. Although the sample was

demographically diverse, it was predominantly White and limited to English-speaking individuals with reliable internet access, potentially restricting generalizability to more diverse or digitally marginalized populations. Future research should intentionally oversample underrepresented groups, including racial and ethnic minorities, non-English speakers, and individuals with limited digital access, to ensure equity in TIC design and evaluation [78]. Second, despite multiple attempts to recruit through the MyChart patient portal, only five participants were enrolled via this channel, limiting meaningful comparisons across recruitment sources and underrepresenting clinical populations. Third, while the regression results highlight meaningful associations between trauma-informed perception and both chatbot design features and user characteristics, the high prevalence of positive outcomes introduces ceiling effects that may inflate odds ratios and reduce estimate stability for rare subgroups. Accordingly, these findings should be interpreted as identifying salient patterns and hypotheses for future validation in more balanced and clinically diverse samples. Finally, although the study relied primarily on quantitative survey data, which provides breadth, it offers limited depth regarding lived experiences. To strengthen data triangulation, the research team is conducting complementary semi-structured interviews as part of the broader project.

In conclusion, our study establishes a rigorous empirical foundation for understanding how TIC principles manifest in mental health chatbot interactions. By integrating descriptive analysis, factor modeling, predictive regression, and qualitative insights, we demonstrate that TIC is both measurable and impactful, with trust, empowerment, and peer support emerging as central drivers of user perceptions alongside key individual characteristics. To advance trauma-informed digital mental health, developers should prioritize hybrid models that integrate peer networks and human–bot collaboration, tailor functionality to diverse user profiles, and systematically evaluate emerging TIC dimensions. Future longitudinal and multi-platform research will be critical for validating and extending these findings, ultimately guiding the design of more inclusive, effective, and psychologically safe chatbot interventions. As artificial intelligence becomes increasingly embedded in mental health service delivery, the ethical imperative to center trauma survivors' voices, protect their safety, and honor their agency has never been more urgent [79]. This study represents an essential step toward realizing that vision.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://gitlab.com/E190153/user-perceptions-of-trauma-informed-chatbots.

Ethics statement

The studies involving humans were approved by the Parkview Health Institutional Review Board. The studies were conducted in accordance with local legislation and institutional requirements. Written informed consent for participation in this study was obtained directly from all participants prior to their involvement in the study.

Author contributions

XL: Formal analysis, Methodology, Visualization, Writing – original draft. SG: Conceptualization, Methodology, Writing – review & editing. RP: Data curation, Writing – review & editing. JP: Writing – review & editing. CW: Conceptualization, Writing – review & editing. FN: Conceptualization, Data curation, Methodology, Writing – review & editing.

Funding

The author(s) declared financial support was received for this work and/or its publication. This work was conducted as part of a study supported by the National Science Foundation Award No. 2348691.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1797681/full#supplementary-material

References

1. Kulkarni P, Mahabaleshwarkar A, Kulkarni M, Sirsikar N, Gadgil KD. Conversational AI: an overview of methodologies, applications & future scope. In: *Proceedings of the International Conference on Computing, Communication, Control and Automation (ICCUBEA)*. Pune, India: IEEE (2019). p. 1–6. doi: 10.1109/ICCUBEA47591.2019.9129347

2. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR Ment Health*. (2017) 4:e19. doi: 10.2196/mental.7785

3. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. *JMIR Ment Health*. (2019) 6: e14166. doi: 10.2196/14166

4. Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. *J Med Internet Res*. (2020) 22:e16021. doi: 10.2196/16021

5. Han Q, Zhao C. Unleashing the potential of chatbots in mental health: bibliometric analysis. *Front Psychiatry*. (2025) 16:1494355. doi: 10.3389/fpsyt.2025.1494355

6. Li H, Zhang R, Lee YC, Kraut RE, Mohr DC. Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being. *NPJ Digit Med*. (2023) 6:1–14. doi: 10.1038/s41746-023-00979-5

7. Meheli S, Sinha C, Kadaba M. Understanding people with chronic pain who use a cognitive behavioral therapy–based artificial intelligence mental health app (Wysa): mixed methods retrospective observational study. *JMIR Hum Factors*. (2022) 9: e35671. doi: 10.2196/35671

8. Haque R, Rubya S. An overview of chatbot-based mobile mental health applications: insights from app description and user reviews. *JMIR Mhealth Uhealth*. (2023) 11:e44838. doi: 10.2196/44838

9. Lau N, O'Daffer A, Colt S, Yi-Frazier JP, Palermo TM, McCauley E, et al. Android and iphone mobile apps for psychosocial wellness and stress management: systematic search in app stores and literature review. *JMIR Mhealth Uhealth*. (2020) 8:e17798. doi: 10.2196/17798

10. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *Can J Psychiatry*. (2019) 64:456–464. doi: 10.1177/0706743719828977

11. Farzan M, Ebrahimi H, Pourali M, Sabeti F. Artificial intelligence-powered cognitive behavioral therapy chatbots, a systematic review. *Iran J Psychiatry*. (2025) 20:102–10. doi: 10.18502/ijps.v20i1.17395

12. Gaffney H, Mansell W, Edwards R, Wright J. Manage your life online (MYLO): a pilot trial of a conversational computer-based intervention for problem solving in a student sample. *Behav Cogn Psychother*. (2014) 42:731–46. doi: 10.1017/S135246581300060X

13. Denecke K, Gabarron E. The ethical aspects of integrating sentiment and emotion analysis in chatbots for depression intervention. *Front Psychiatry*. (2024) 15:1462083. doi: 10.3389/fpsyt.2024.1462083

14. Hipgrave L, Goldie J, Dennis S, Coleman A. Balancing risks and benefits: clinicians' perspectives on the use of generative AI chatbots in mental healthcare. *Front Digit Health*. (2025) 7:1606291. doi: 10.3389/fdgth.2025.1606291

15. Lawrence HR, Schneider RA, Rubin SB, Matarić MJ, McDuff DJ, Jones Bell M. The opportunities and risks of large language models in mental health. *JMIR Ment Health*. (2024) 11:e59479. doi: 10.2196/59479

16. Rahsepar Meadi M, Sillekens T, Metselaar S, van Balkom A, Bernstein J, Batelaan N. Exploring the ethical challenges of conversational AI in mental health care: scoping review. *JMIR Ment Health*. (2025) 12:e60432. doi: 10.2196/60432

17. Abd-Alrazaq A, Al-Mudawar M, Alhuwail D, Al-Kuwari S, Nashwan AJ, Shah Z. Large language models in mental health care: a scoping review. *npj Digit Med*. (2023) 6:167. doi: 10.1038/s41746-023-00898-7

18. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. (2023) 29:1930–40. doi: 10.1038/s41591-023-02448-8

19. De Freitas J, Uğuralp AK, Oğuz-Uğuralp Z, Puntoni S. Chatbots and mental health: insights into the safety of generative AI. *J Consum Psychol*. (2024) 34:481–91. doi: 10.1002/jcpy.1393

20. Scholich T, Barr M, Wiltsey Stirman S, Raj S. A comparison of responses from human therapists and large language model–based chatbots to assess therapeutic communication: mixed methods study. *JMIR Ment Health*. (2025) 12:e69709. doi: 10.2196/69709

21. Brown JEH, Halpern J. AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM Ment Health*. (2021) 1:100017. doi: 10.1016/j.ssmmh.2021.100017

22. Roose K. Data from: Can AI be blamed for a teen's suicide? *The New York Times*. (2024). Available online at: https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html (Accessed October 23).

23. Xiang C. Data from: "He would still be here": man dies by suicide after talking with AI chatbot, widow says. *Vice* (2023). https://www.vice.com/en/article/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says (Accessed March 30).

24. Eggleston M, Jones EP, Khan N, Romani WA, McQuillan K, Otero J, et al. A scoping review of trauma-informed care principles applied in design and technology. *Digit Health*. (2025) 11. doi: 10.1177/20552076251360925

25. Mazuz K, Yamazaki R. Trauma-informed care approach in developing companion robots: a preliminary observational study. *Front. Rob AI*. (2025) 12:1476063. doi: 10.3389/frobt.2025.1476063

26. Substance Abuse and Mental Health Services Administration. *Trauma-informed care in behavioral health services*. No. HHS Publication No. (SMA) 13-4801 in Treatment Improvement Protocol (TIP) Series 57. Rockville, MD: Substance Abuse and Mental Health Services Administration (2014b). (Accessed January 20, 2026).

27. Substance Abuse and Mental Health Services Administration. *SAMHSA's concept of trauma and guidance for a trauma-informed approach* (2014a). Tech. rep., SAMHSA, Rockville, MD. HHS Publication No. SMA14-4884.

28. Abdulai A-F. Is generative AI increasing the risk for technology-mediated trauma among vulnerable populations? *Nurs Inq*. (2025) 32:e12686. doi: 10.1111/nin.12686

29. Chen JX, McDonald A, Zou X, Tseng E, Roundy KA, Tamersoy A, et al. Trauma-informed computing: towards safer technology experiences for all. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM (2022). p. 1–15. doi: 10.1145/3491102.3517475

30. Felitti VJ, Anda RF, Nordenberg D, Williamson DF, Spitz AM, Edwards V, et al. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: the adverse childhood experiences (ACE) study. *Am J Prev Med*. (1998) 56:774–86. doi: 10.1016/S0749-3797(98)00017-8

31. Gerber MR, Elisseou S, Sager ZS, Keith JA. Trauma-informed telehealth in the COVID-19 era and beyond. *Fed Pract*. (2020) 37:302–8. doi: 10.1016/S0749-3797(98)00017-8

32. Muskett C. Trauma-informed care in inpatient mental health settings: a review of the literature. *Int J Ment Health Nurs*. (2014) 23:51–9. doi: 10.1111/inm.12012

33. Randazzo C, Scott CF, Bellini R, Ammari T, Devito MA, Semaan B, et al. Trauma-informed design: a collaborative approach to building safer online spaces. In: *Proceedings Companion Publication of the Conference on Computer Supported Cooperative Work and Social Computing*. ACM (2023). p. 203–7. doi: 10.1145/3584931.3611277

34. Substance Abuse and Mental Health Services Administration. Data from: Trauma-informed approaches and programs (2024). Available online at: https://www.samhsa.gov/mental-health/trauma-violence/trauma-informed-approaches-programs (Accessed December 3, 2024).

35. Lee HS, Wright C, Ferranto J, Buttimer J, Palmer CE, Welchman A, et al. Artificial intelligence conversational agents in mental health: patients see potential, but prefer humans in the loop. *Front Psychiatry*. (2025) 15:1505024. doi: 10.3389/fpsyt.2024.1505024

36. Schäfer LM, Krause T, Köhler S. Exploring user characteristics, motives, and expectations and the therapeutic alliance in the mental health conversational AI Clare®: a baseline study. *Front Digit Health*. (2025) 7:1576135. doi: 10.3389/fdgth.2025.1576135

37. Zou X, Na Y, Lai K, Liu G. Unpacking public resistance to health chatbots: a parallel mediation analysis. *Front Psychol*. (2024) 15:1276968. doi: 10.3389/fpsyg.2024.1276968

38. Kollig F, Voelker K, Ryan E, Pfafman R, Nova FF. Current landscape of mental health conversational agents from a trauma-informed care lens: scoping review (2025). *JMIR Preprints* [Accepted preprint].

39. Kostenius C, Lindström F, Potts C, Pekkari N. Young peoples' reflections about using a chatbot to promote their mental wellbeing in northern periphery areas—a qualitative study. *Int J Circumpolar Health*. (2024) 83. doi: 10.1080/22423982.2024.2369349

40. National Coalition for Sexual Health. Data from: Trauma-informed care guide (2023). Available online at: https://www.nationalcoalitionforsexualhealth.org/tools/for-healthcare-providers/trauma-informed-care-guide (Accessed March 31, 2026).

41. Scott CF, Marcu G, Anderson RE, Newman MW, Schoenebeck S. Trauma-informed social media: towards solutions for reducing and healing online harm. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery (2023). CHI '23. doi: 10.1145/3544548.3581512

42. Substance Abuse and Mental Health Services Administration. *Practical guide for implementing a trauma-informed approach* (2023). Tech. rep., SAMHSA, Rockville, MD. HHS Publication No. PEP23-06-05-005.

43. Clickworker GmbH. Data from: Clickworker: crowdsourcing marketplace (2026). Available online at: https://www.clickworker.com/ (Accessed January 14, 2026).

44. Gnewuch U, Morana S, Hinz O, Kellner R, Maedche A. More than a bot? The impact of disclosing human involvement on customer interactions with hybrid service agents. *Inf Syst Res*. (2024) 35:936–55. doi: 10.1287/isre.2022.0152

45. Braun V, Clarke V. Reflecting on reflexive thematic analysis. *Qual Res Sport Exerc Health*. (2019) 11:589–97. doi: 10.1080/2159676X.2019.1628806

46. Luger E, Sellen A. "Like having a really bad PA": the gulf between user expectation and experience of conversational agents. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. (2016a). p. 5286–97. doi: 10.1145/2858036.2858288

47. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. *Epidemiol Psychiatr Sci*. (2016) 25:113–22. doi: 10.1017/S2045796015001067

48. Seeger A-M, Pfeiffer J, Heinzl A. Texting with humanlike conversational agents: designing for anthropomorphism. *J Assoc Inf Syst*. (2021) 22:931–67. doi: 10.17705/1jais.00685

49. Calvo R, Peters D, Vold K, Ryan R. Supporting human autonomy in AI systems: a framework for ethical enquiry. In: *Philosophical Studies Series*. Cham: Springer (2020). p. 31–54. doi: 10.1007/978-3-030-50585-1_2

50. Ryan RM, Deci EL. *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. New York, NY: Guilford Press (2017).

51. Bickmore TW, Trinh H, Olafsson S, O'Leary TK, Asadi R, Rickles NM, et al. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google assistant. *J Med Internet Res*. (2018) 20:e11510. doi: 10.2196/11510

52. He Y, Yang L, Zhu X, Wu B, Zhang S, Qian C, et al. Mental health chatbot for young adults with depressive symptoms during the COVID-19 pandemic: single-blind, three-arm randomized controlled trial. *J Med Internet Res*. (2022) 24:e40719. doi: 10.2196/40719

53. Potts C, Lindström F, Bond R, Mulvenna M, Booth F, Ennis E, et al. A multilingual digital mental health and wellbeing chatbot (chatpal): pre-post multicenter intervention study. *J Med Internet Res*. (2023) 25:e43051. doi: 10.2196/43051

54. Moilanen J, van Berkel N, Visuri A, Gadiraju U, van der Maden W, Hosio S. Supporting mental health self-care discovery through a chatbot. *Front Digit Health*. (2023) 5:1034724. doi: 10.3389/fdgth.2023.1034724

55. Tremain H, McEnery C, Fletcher K, Murray G. The therapeutic alliance in digital mental health interventions for serious mental illnesses: narrative review. *JMIR Ment Health*. (2020) 7:e17204. doi: 10.2196/17204

56. Beilharz F, Sukunesan S, Rossell SL, Kulkarni J, Sharp G. Development of a positive body image chatbot (kit) with young people and parents/carers: qualitative focus group study. *J Med Internet Res*. (2021) 23:e27807. doi: 10.2196/27807

57. Ahlers K, Gabrielli J. A systematic review of internet communication technology use by youth in foster care. *Child Adolesc Social Work J*. (2021) 39:375–90.

58. Brown D. Childhood experiences, growing up "in care," and trust: a quantitative analysis. *Child Youth Serv Rev*. (2022) 140:106584. doi: 10.1016/j.childyouth.2022.106584

59. Folk JB, Gill AM, Ordorica C, Rodriguez C, Holloway IW, Meza RD, et al. An ecodevelopmental framework for engaging diverse youth in foster care and their families into technology-based family intervention research trials. *Front Digit Health*. (2022) 4:866139. doi: 10.1016/S0749-3797(98)00017-8

60. Patmisari E, Jones M, Kang J, Metelak R, Ordorica C, Tolou-Shams M. Reflections of foster youth engaging in the co-design of digital mental health technology: duoethnography study. *JMIR Ment Health*. (2025) 9:e53231. doi: 10.1016/S0749-3797(98)00017-8

61. Ahrens KR, Garnett MF, Dugan J, McCarthy M, Doty JG, Schultz S, et al. Intervention development to improve foster youth mental health by targeting coping self-efficacy and help-seeking. *J Child Fam Stud*. (2023) 144:106753.

62. Bennett K, Courtney J, Decker K, Duppong Hurley K, Pfeiffer H. Trust is essential: identifying trust building techniques from youth providers across the service array. *Child Youth Serv Rev*. (2020) 116:105320. doi: 10.1016/j.childyouth.2020.105320

63. Singer ER, Berzin SC. Former foster system youth: perspectives on transitional supports and programs. *J Soc Serv Res*. (2020) 48:287–305.

64. Leo HP, Folk JB, Rodriguez C, Tolou-Shams M. Implementation considerations for family-based telehealth interventions for youth in foster care: focus group study with child welfare system professionals. *JMIR Form Res*. (2023) 7:e45905. doi: 10.2196/45905

65. Advokids. Data from: Attachment (2024). Available online at: https://advokids.org/childhood-mental-health/attachment/ (Accessed December 13, 2025).

66. Bovenschen I, Lang K, Zimmermann J, Förthner J, Nowacki K, Roland I, et al. Foster children's attachment security in the first year after placement: a longitudinal study of predictors. *Early Child Res Q*. (2016) 36:269–80. doi: 10.1016/S0749-3797(98)00017-8

67. Considering Adoption. Data from: Attachment in fostering and adoption (2020). Available online at: https://consideringadoption.com/parenting-after-foster-care-adoption/rad-and-attachment/ (Accessed December 13, 2025).

68. Indiana Adoption Lawyer. Data from: Summary of attachment and bonding in the foster and adopted child (2024). Available online at: https://www.hoosierfamilylawyer.com/foster-children-adoption-book/summary-attachment-and-bonding-foster-and-adopted-child (Accessed December 14, 2025).

69. Research in Practice. Data from: Attachment theory and research: fostering and adoption (2024). Available online at: https://fosteringandadoption.rip.org.uk/topics/attachment-theory-research/ (Accessed December 14, 2025).

70. Bickmore TW, Mitchell SE, Jack BW, Paasche-Orlow MK, Pfeifer LM, O'Donnell J. Response to a relational agent by hospital patients with depressive symptoms. *Interact Comput*. (2010) 22:289–98. doi: 10.1016/j.intcom.2009.12.001

71. Bickmore TW, Picard RW. Establishing and maintaining long-term human-computer relationships. *ACM Trans Comput Hum Interact*. (2005) 12:293–327. doi: 10.1145/1067860.1067867

72. Moussavi Y, Haugland BSM, Wergeland GJ, Lehmann S, Mæland S. Experiences from treatment for anxiety and depression among youth in foster care: a qualitative study. *Child Fam Soc Work*. (2025) 30:342–53. doi: 10.1111/cfs.13172

73. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth*. (2018) 6:e12106. doi: 10.2196/12106

74. iFoster. Data from: Foster youth are the most disconnected young people (2024). Available online at: https://www.ifoster.org/bridging-the-digital-divide/ (Accessed December 14, 2025).

75. Følstad A, Nordheim CB, Bjørkli CA. What makes users trust a chatbot for customer service? An exploratory interview study. In: Bodrunova SS. editor. *Internet Science: 5th International Conference, INSCI 2018*. Cham: Springer (2018). p. 194–208. Lecture Notes in Computer Science; Vol. 11193. doi: 10.1007/978-3-030-01437-7_16

76. Luger E, Sellen A. "Like having a really bad PA": the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM (2016b). CHI '16. p. 5286–97. doi: 10.1145/2858036.2858288

77. Volpato R, DeBruine L, Stumpf S. Trusting emotional support from generative artificial intelligence: a conceptual review. *Comput Hum Behav Artif Hum*. (2025) 5:100195. doi: 10.1016/j.chbah.2025.100195

78. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: How informatics interventions can worsen inequality. *J Am Med Inform Assoc*. (2018) 25:1080–8. doi: 10.1093/jamia/ocy052

79. Martínez-Martín N, Luo Z, Kaushal A, Adeli E, Haque A, Kelly S, et al. Ethical issues for direct-to-consumer digital psychotherapy apps: addressing accountability, data protection, and consent. *JMIR Ment Health*. (2020) 7:e15423. doi: 10.2196/15423