"I Just Don't Want to Be a Burden": Investigating AI Chatbot Emotional Support Among People with Suicidal Ideation and Other Mental Health Concerns

VASUDHA VARADARAJAN, Language Technologies Institute, Carnegie Mellon University, USA
NATHAN S FISHBEIN, Department of Psychology, Fordham University, USA
ANDREW AQUILINA, Department of Informatics and Networked Systems, University of Pittsburgh, USA
JORDAN TAYLOR, School of Information, University of Michigan, USA
HONG SHEN, Human-Computer Interaction Institute, Carnegie Mellon University, USA
MAARTEN SAP, Language Technologies Institute, Carnegie Mellon University, USA and Allen Institute for AI, USA

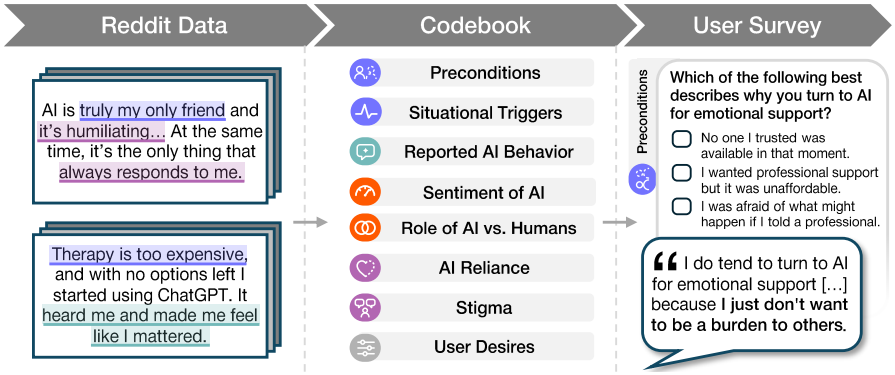


Fig. 1. **Overview of the sequential mixed-methods study design.** We first analyzed Reddit posts describing AI use and mental health-related distress and inductively developed a codebook. We then translated these constructs into structured survey items to assess whether the patterns identified in Reddit appeared in a broader sample of people who had used AI for emotional support and reported psychiatric diagnoses or high-acuity mental health problems. Reddit excerpts shown in the figure are paraphrased.

People experiencing high-acuity mental health concerns are turning to AI chatbots for emotional support, but little is known about what drives this use, their experiences, and their preferences. In a sequential mixed-methods study, we inductively analyzed Reddit posts and used the findings to develop a survey. We surveyed 202 adults with recent mental health concerns who had used AI for emotional support. Most participants were already receiving clinical care and reported turning to AI, primarily because they did not want to burden others, rather than being driven by barriers to care. Benefits were broadly shared, but participants reporting past-year

Authors' Contact Information: Vasudha Varadarajan, vvaradar@andrew.cmu.edu, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA; Nathan S Fishbein, nfishbein@fordham.edu, Department of Psychology, Fordham University, Manhattan, NY, USA; Andrew Aquilina, andrew.aquilina@pitt.edu, Department of Informatics and Networked Systems, University of Pittsburgh, Pittsburgh, PA, USA; Jordan Taylor, jordata@umich.edu, School of Information, University of Michigan, Ann Arbor, MI, USA; Hong Shen, hongs@andrew.cmu.edu, Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA; Maarten Sap, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA and Allen Institute for AI, Seattle, WA, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

suicidal ideation described more harm-facilitating AI behavior and reliance than those without. Overall, participants ranked honesty above warmth, and preference for crisis escalation declined as clinical acuity rose. These findings motivate AI support that addresses perceived costs of disclosure, evaluates benefits alongside harms, and widens feasible paths to human connection.

CCS Concepts: • **Human-centered computing** → *Empirical studies in HCI*; **Social networking sites**; **User centered design**; • **Applied computing** → **Psychology**; • **General and reference** → *Evaluation*; *Validation.*

Additional Key Words and Phrases: Mental Health, AI use, Suicidal Ideation, Social Media

ACM Reference Format:

Vasudha Varadarajan, Nathan S Fishbein, Andrew Aquilina, Jordan Taylor, Hong Shen, and Maarten Sap. 2026. *"I Just Don't Want to Be a Burden":* Investigating AI Chatbot Emotional Support Among People with Suicidal Ideation and Other Mental Health Concerns. 1, 1 (September 2026), 23 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Content Warning: This paper discusses mental health topics that may be distressing to some readers.

1 Introduction

In the United States, suicide rates rose 37% from 2000 to 2018 and have remained near record highs since then [18, 19]. People with mental health problems, including suicidal ideation, often struggle to receive timely support. Formal care is difficult to reach: nearly half of adults with a mental illness did not receive treatment in 2024 [11], and over 163 million Americans live in areas with a shortage of mental health providers [8]. Though informal support from friends, family, or peers can protect against psychological distress [13], this type of support is increasingly rare, and the Surgeon General and WHO both recently declared loneliness a public health priority [34, 49]. Critically, how people perceive their relationships may shape both suicidal desire and their willingness to seek support. The interpersonal theory of suicide identifies perceived burdensomeness, or the belief that others would be better off without oneself, as a contributor to suicidal desire [45]. Prior work support an associations between perceived burdensomenes and suicidal ideation [12]. These issues raise questions about where people can and do turn during acute distress and how their perceptions of their relationships shape whether available support feels possible to seek.

People are increasingly turning to generative AI chatbots for emotional support, using systems such as ChatGPT to disclose distress, seek advice, and work through personal difficulties [24, 41]. Prior research suggests that these systems can feel accessible, private, and nonjudgmental, including for people navigating social anxiety [47]. Beyond general-purpose chatbots, AI companions can become relational figures in users' lives, offering a sense of connection while also raising questions about dependency, well-being, and the appropriate role of AI in mental health support [51].

Despite growing attention to generative AI in mental health, we know comparatively little about its use by people experiencing mental health concerns that may involve risks of serious harm, including suicidal ideation, delusions, and mania. Such use is not confined to specialized clinical tools: acute episodes unfold through the same general-purpose systems used for everyday, lower-stakes emotional support [27]. Although OpenAI estimates that only 0.15% of weekly active ChatGPT users have conversations containing explicit indicators of potential suicidal planning or intent, this proportion represents a substantial population at platform scale [35]. Studies of chatlogs have documented how chatbots may amplify delusional beliefs when users experience suicidal ideation [4, 32]. Further, existing user studies evaluating support-seeking on social media do not distinguish between general emotional support and support for acute, high-risk mental health conditions [24].

Existing LLM safety research has largely taken a complementary approach, where benchmarks typically evaluate model responses against expert-defined safety criteria [7, 48]. These evaluations are critical, but they often treat crises

as discrete and readily identifiable exchanges, potentially assuming that users recognize that they are in crisis, seek a clinical form of intervention, and can act on referrals to formal care. User-centered studies have examined the social contexts that shape help-seeking, including the design and availability of mental health resources, as well as users' attitudes toward AI during crises – often developed through in-depth interviews with smaller populations [37, 41, 50]. We build on this foundation by examining the experience more holistically, seeking to develop a broader account of **what leads people with psychiatric diagnoses or high-acuity mental health problems, including suicidal ideation, to use AI for support, how those exchanges unfold, and how users perceive and respond to them afterward**. Specifically, we ask:

- **RQ1**: What situations and circumstances lead people in this population to seek support from chatbots?
- **RQ2**: How does this group describe their experiences of using chatbots for support?
- **RQ3**: What do people in this group need and want chatbots to do, and not do, when providing emotional or crisis-related support?

To address these questions, we conducted a sequential mixed-methods study that uses exploratory qualitative analysis to identify patterns of AI use among those reporting mental health concerns and a subsequent survey to assess whether and how those patterns generalize across a broader population of users.

We first analyze Reddit discussions in which users describe mental health concerns and turning to generative AI for emotional support. Using a hybrid inductive-deductive analytic approach, we develop a codebook capturing the circumstances preceding AI use, the forms of support and interaction that emerge, and users' subsequent appraisals of AI. This exploratory analysis provides a detailed and user-described account of heterogeneous experiences of AI use among people reporting mental health concerns across Reddit communities, while generating empirically grounded patterns and questions for further examination. We then use these findings to inform the design of a targeted user survey of Prolific participants who report lived experience of psychiatric diagnoses or high-acuity mental health problems, including suicidal ideation, and prior use of AI for support during distress. The survey assesses how patterns identified in the Reddit data resonate across a broader set of experiences and enables more direct examination of users' needs, expectations, current use, and changing relationships with AI across stages of seeking support.

Drawing on both studies, we discuss implications for mental health practitioners, LLM developers, AI safety and trust teams, and platform-level safeguards for AI use among our target population, as well as opportunities to distinguish use during acute crises from lower-stakes forms of emotional support. This work makes three contributions:

- A codebook describing an exploratory account of AI use in the setting of self-reported psychiatric diagnoses or high-acuity mental health problems, connecting the circumstances that lead to use, the interactions that unfold, longer-term reliance, and users' desired and undesired AI behaviors.
- An empirical study of our account of AI use among this high-risk population, evaluated to parse the conditions, experiences and desires of participants with different clinical concerns.
- Implications for LLM developers and clinicians working with patients who use AI for emotional support.

2 Related Work

2.1 Specialized Chatbots for Mental Healthcare

The potential of conversational systems to expand access to mental health support has long been studied, with earlier systems designed for specific interventions. For instance, Woebot delivered cognitive behavioral therapy content in a randomized trial with young adults, while Wysa combined conversational support with structured exercises [16, 23].

This literature found promising evidence but substantial variation in system design, study quality, and outcome measurement [17]. More recently, specialized systems have incorporated generative AI. Therabot, an expert-fine-tuned generative chatbot, was evaluated in a randomized trial for mental-health treatment [20]. This work demonstrates the capacity of conversational agents to deliver psychological interventions under controlled conditions.

However, this research leaves an important form of support seeking unexamined: how and why people seek emotional support from general-purpose chatbots not designed as mental-health interventions. Mental health exchanges occur with chatbots used for everyday questions, writing, or other tasks, without users first selecting a dedicated mental health intervention [24, 42]. Studying use outside research settings is particularly important because self-directed digital mental health interventions show a substantial research-to-practice gap, with real-world engagement falling well below trial engagement [6]. Understanding these exchanges requires examining what makes a general-purpose system an accessible and acceptable place to disclose distress.

2.2 General-Purpose Chatbots Used for Mental Healthcare

The broad conversational appeal of general-purpose chatbots allows users to move between practical questions and emotional concerns within the same interface. For an existing user, emotional disclosure can therefore begin within a familiar tool, without a separate search for a mental-health service. Research describes users valuing immediate availability, flexible dialogue, and a perceived absence of judgment or interpersonal obligations [24, 41]. Some participants in Siddals et al. [41] also contrasted generative chatbots favourably with the scripted interactions of earlier mental-health apps, pointing to improved conversational flexibility and personal connection. These findings help explain how support-seeking can emerge through an accessible conversational tool.

Analyses of Reddit discussions describe chatbots providing validation, reflection, information, and concrete next steps, while also raising concerns about privacy, incorrect advice, and excessive agreement [24]. Interview research likewise documents emotionally meaningful support alongside limitations in memory, guidance, and relational understanding [41]. To characterize these tensions, Song et al. [42] introduce *therapeutic alignment*: aligning AI used for mental-health support with therapeutic values and users' broader needs. Their account emphasizes how users actively construct supportive roles for chatbots, including roles that complement human care. These needs are heterogeneous: people with differing levels of social anxiety place different weight on non-judgmental responding, emotional understanding, trust, and reliability [47], while interviews with people living with mental-health conditions describe situational use and boundaries around what chatbots can offer during more complex distress [38].

This work situates AI use within both gaps in care and the perceived social costs of disclosure. Technology-mediated help-seeking research similarly shows that timing, availability, and the organization of services shape which forms of support people can use [37]. However, the same accessibility that makes general-purpose chatbots appealing for everyday emotional support can also place them in roles with greater clinical stakes. The roles users assign to these systems may conflict with clinical recommendations, raising questions about how their responses should be evaluated in these moments.

2.3 Clinical Recommendations for and Evaluations of Generative AI for Mental Health

Normative research on generative AI and mental health has largely recommended that AI aid practitioners in providing care but not independently deliver treatment or function as a crisis resource. Some of this work proposes a staged model, in which AI advances slowly from simple tasks, such as psychoeducation, to more complex ones, such as treatment planning [43]. Autonomous care remains a distant, high-risk possibility [43]. This cautious approach reflects concerns

that AI cannot assume the obligations and accountability of a therapist, particularly in risky interactions with suicidal users [26]. Some policymakers and professional groups have adopted similar perspectives [5, 22], and the American Psychological Association has called general-purpose chatbots “unpredictable" in crisis management [3]. Despite these recommendations, many people continue to use general-purpose chatbots during acute distress [35], leading experts to call for research assessing how these systems respond to mental-health concerns [21].

Though clinically grounded evaluations of general purpose chatbots are proliferating, none have been shown to predict safety or outcomes among high-risk real-world users. These evaluations typically examine model behavior under simulated conditions, using responses to vignettes or simulated users to compare AI behavior with clinician responses or against clinician-developed rubrics [14, 30, 31]. Performance varies by assessment: models respond appropriately to suicide-related questions at very high and very low risk levels but diverge from expert judgment at intermediate levels [29]; they judge the appropriateness of responses to suicidal ideation with varying accuracy, at times exceeding trained clinicians [30], and underperform clinicians in safely responding to content with delusions [31]. However, many assessments evaluate single-turn interactions and may therefore miss harms that accumulate across multi-turn conversations [33]. While benchmarks designed to evaluate risk over multiple turns are emerging [7, 9], they remain simulated. Clinical trials provide another means of assessing outcomes related to chatbot behavior over multi-turn interactions. While they report reduced symptoms in mild-moderate clinical samples, they tend to exclude those with acute or high-risk conditions and use specialized AI [20, 40]. Their applicability to general-purpose systems used during crises is therefore limited.

User-centered research provides complementary evidence about the circumstances surrounding use and how users perceive the support they receive. These studies use retrospective interviews, user accounts, and donated chat logs to study experiences with AI mental health support [32, 38, 41, 42]. A small portion of these studies report on crisis use and delusional spirals [2, 32]. These studies report subjective experiences and interpretations that AI-behavior-focused research is unable to evaluate. However, the contexts that prompt people to seek emotional support from AI, how people describe and evaluate that support, and their preferences for AI behavior during distress remain incompletely connected across this literature. Our study examines these dimensions together among people with recent mental health concerns, including suicidal ideation.

3 Study 1: Reddit study

We first grounded our investigation in existing Reddit discussions to understand how people were already describing AI use for emotional support in their own terms. Reddit’s pseudonymous, discussion-oriented communities offer a relatively low-barrier setting for users to share perceived benefits as well as concerns, and recent AI-focused communities have produced substantial discourse around AI therapy, companionship, and emotional reliance [1, 36].

3.1 Data Collection

Subreddit and Post Identification. Inspired by Shen et al. [39], we used an iterative, search-based procedure to identify Reddit communities where users discussed AI use alongside severe emotional distress or suicide-related concerns. We initially searched Reddit using `AI use`, `mental health`, `suicidal`, and `emotional support`, reviewed resulting posts and subreddits for relevance, and iteratively refined the query set using terminology from relevant results. The final terms were: `mental health`, `emotional support`, `loneliness`, `depression`, `psychosis`, `anxiety`, `therapy`, `AI use`, `AI for mental health`, `therapist`, `talk to AI`, `suicidal`, and `AI companion`.

We continued until thematic saturation, when additional searches identified no new relevant subreddits. This yielded 34 subreddits: EDAnonymous, addiction, alcoholism, adhd, anxiety, autism, bipolarreddit, bpd, depression, healthanxiety, lonely, ptsd, schizophrenia, SuicideWatch, socialanxiety, suicidewatch, mentalhealth, COVID19_support, mentalillness, therapy, therapists, raisedbynarcissists, askatherapist, BeyondtheAIAssistant, TherapyGPT, AI_For_Therapy, MyBoyfriendIsAI, AIRelationships, BeyondThePromptAI, ChatGPT, antiai, AITherapySpace, ClaudeAI, and CPTSD. We removed duplicate subreddit names from the final list.

Data Collection and Filtering. We collected 3,342,335 posts published in these communities between January 1, 2023, and January 31, 2026, to capture AI use specifically in the post-LLM era. Sampling across communities captured diverse mental-health experiences, help-seeking contexts, and relationships with AI, rather than a narrower subset of Reddit users or experiences.

To identify posts discussing both AI use and elevated mental-health or suicide-related distress, we adapted the archetype-based suicide-risk inference method of Varadarajan et al. [46]. We curated separate prototype sets for (1) AI use and (2) severe emotional distress or suicide-related risk. Suicide-risk prototypes were adapted from Varadarajan et al. [46], while AI-use prototypes were curated by the authors (see the supplementary material).

We embedded prototypes and post sentences using `sentence-transformers/all-MiniLM-L6-v2`. For each construct, a post's score was the maximum cosine similarity between any of its sentences and the relevant prototype set. Prespecified thresholds of 0.30 for AI use and 0.50 for suicide-related risk yielded approximately 2,008 posts meeting both criteria.[1] The resulting candidate set contained posts referencing AI use in the context of mental-health concerns, including severe emotional distress and suicide-related risk.

3.2 Codebook Development and Validation

To characterize how Reddit users described their use of conversational AI in relation to mental-health support, we conducted an inductive qualitative analysis of 100 of the filtered Reddit posts. Rather than treating AI use as an isolated behavior, we examined it as situated within the posters' broader accounts of distress, available support, care access, interpersonal relationships, and everyday coping. Our analytic focus was therefore the portion of each post in which the author mentioned using, or considering using an AI system, alongside the surrounding text needed to interpret the role that AI played in that account.

Two authors independently reviewed and conducted the analysis: one with training in clinical psychology and mental-health research, and one with expertise in computer science. This interdisciplinary pairing supported attention both to the psychosocial context of posters' narratives and to the particular affordances, risks, and interactional roles attributed to AI systems. We used a hybrid inductive and deductive approach [15], allowing codes to emerge from participants' descriptions rather than applying a predetermined framework of mental-health needs, technology use, or AI harms.

Inductive Analysis. We used a two-stage qualitative coding process. First, two authors independently open-coded the same 100 Reddit posts, focusing on passages in which posters described using, considering, or reflecting on AI and the surrounding context needed to interpret AI's role. Researchers generated descriptive codes inductively, without a predetermined codebook. These codes captured explicit reasons for AI use (e.g., therapy cost or unavailability), as well as associated conditions, experiences, and perceived consequences, including immediate emotional release, feeling understood, frequent use, and concern about reliance on AI.

[1] We selected thresholds by reviewing the 20 lowest-scoring posts retained under candidate cutoffs and choosing values for which at least 50% of borderline posts mentioned the corresponding construct.

The researchers then discussed candidate codes, and one synthesized an initial codebook by identifying recurring concepts, consolidating overlaps, and drafting definitions and inclusion criteria. The second researcher reviewed and revised this codebook, proposing additional codes and refinements to code boundaries. Disagreements were resolved through discussion, resulting in a shared codebook representing both researchers' interpretations of the initial 100 posts.

Deductive Analysis. Using the shared codebook, each researcher independently coded an additional 30 posts, applying established codes while flagging unclear or insufficiently represented cases. Posts could receive multiple codes, reflecting intersecting motivations, constraints, use patterns, and perceived outcomes. For example, a poster might describe using an AI chatbot because of loneliness and financial barriers to therapy, find it immediately comforting, and also express concern that use had become difficult to control. We therefore treated codes as non-mutually exclusive.

The researchers reviewed ambiguous cases from this stage to determine whether code definitions required clarification or the codebook needed expansion. The final hierarchical codebook comprised 8 Level 1 codes and 27 nested Level 2 codes.

Finally, three authors (one new to the codebook), independently annotated 100 posts by selecting spans and assigning codes. To assess reliability, we matched spans across annotator pairs using longest common subsequence (LCS) token overlap, retaining mutually best-matching spans with LCS ≥ 3 tokens. This produced 216 matched-span units: 113 annotated by all three annotators and 103 by two. Krippendorff's α (nominal) was 0.558 for Level 1 codes and 0.481 for Level 2 codes, indicating low-to-fair agreement given the codebook's fine-grained structure.

The final codebook was developed from qualitative analysis of 230 posts across 17 Reddit communities: `depression`, `therapy`, `ChatGPT`, `ptsd`, `mentalillness`, `BipolarReddit`, `mentalhealth`, `antiai`, `CPTSD`, `SuicideWatch`, `therapyGPT`, `therapists`, `raisedbynarcissists`, `autism`, `AIRelationships`, and `MyBoyfriendisAI`.

3.3 Final Codebook

We organized the codebook around eight broad Level 1 categories (A–H), conceptualizing posters' accounts as unfolding along a rough trajectory of AI use for acute emotional distress rather than as a flat, unordered set of themes, shown in Figure 2.

Categories A and B capture the conditions that precede AI use. Category A, structural preconditions, describes standing conditions that make a poster more likely to turn to AI at all, while Category B, situational triggers, captures the immediate circumstance or momentary triggers for a specific instance of use. Category C, reported AI behavior, captures what the AI actually did during the interaction, as described by the poster. Categories D and E capture the immediate aftermath of an interaction. Category D, sentiment and evaluation, tracks how the poster judged the AI's behavior, and Category E, role of AI relative to human support, captures how posters positioned the two forms of support against each other. Categories F and G shift from single interactions to longer-term patterns. Category F, AI–user bond and reliance, covers perceived reliance and relational bonding, while Category G covers societal and self-directed stigma. Finally, Category H, user desires and design critiques, departs from the trajectory of AI use to capture what posters wished AI had done differently, independent of any single interaction. The 27 Level 2 codes, with full definitions, inclusion and exclusion rules, and illustrative post examples, are provided in the supplementary material.

As shown in Figure 2, the eight categories are organized into five broader phases: conditions preceding AI use, use itself, immediate post-use experiences, longer-term effects, and desired AI behaviors. This organization reflects the longitudinal process structure through which people describe using AI for emotional support.

However, individual posts did not necessarily describe every phase of this process. Some posters discussed only the circumstances that led them to consider or use AI, while others focused primarily on desired improvements to AI-based

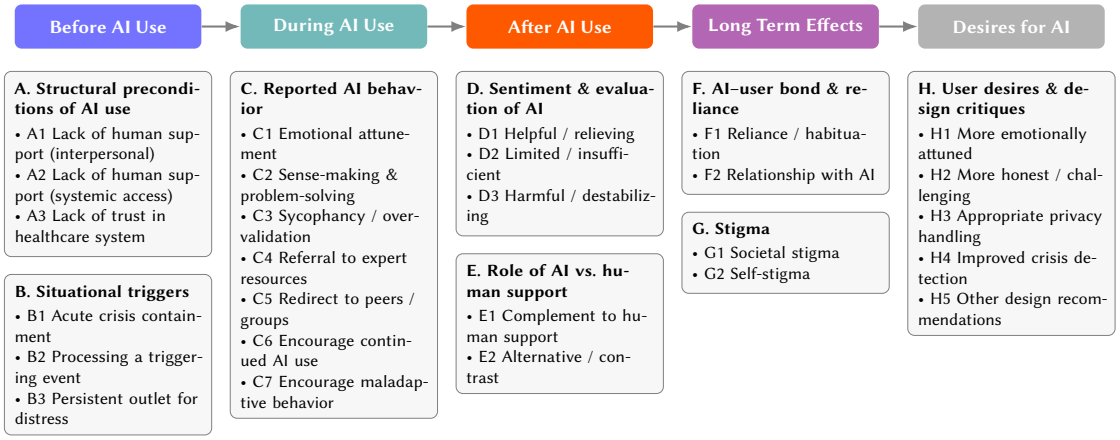


Fig. 2. Organization of the codebook's Level 1 categories (A–H) and their constituent Level 2 codes across the arc of an AI-support interaction: circumstances preceding use (Before AI Use), the interaction itself as reported by the poster (During AI Use), the poster's appraisal of the interaction (After AI Use), patterns of bond, reliance, and stigma accumulating across repeated use (Long Term Effects of AI use), and the poster's stated wishes for how AI support could be improved (Desires for AI). Category and code labels correspond directly to the full codebook provided in the supplementary material.

emotional support. Some others primarily described worsening mental health, with only brief or occasional references to AI as a coping resource. To further examine these dynamics beyond the partial accounts available in individual posts, we developed a user survey based on the codebook. The survey captures the multiple conditions, practices, experiences, and longer-term effects that may shape AI use for emotional support during periods of distress.

4 Study 2: User Survey

While Reddit analysis surfaced heterogeneous and otherwise inaccessible accounts of AI use during emotional distress, individual posts provide partial, situated, and community-shaped accounts. Following our initial research questions: why people with psychiatric diagnoses and other high-acuity mental health problems turn to AI (**RQ1**), how they describe their experiences of using AI for support (**RQ2**), and how users evaluate its benefits and harms (**RQ3**), we conducted a targeted user study based on the developed codebook. We designed a cross-sectional survey of people with recent mental-health concerns who had used an AI chatbot for emotional or mental-health support, translating the qualitative codebook into structured items on users' circumstances, interactions, experiences, and desired AI behavior.

4.1 Recruitment and Eligibility

We recruited participants through Prolific and administered the study in Qualtrics. The eligibility study was shown to adults aged 18 years or older who lived in the United States and reported fluency in English. Participants were eligible for the survey if they reported (1) at least one of the following within the past 12 months: wishing they were dead or would not wake up, thoughts of killing themselves, intentional self-harm without suicidal intent, emergency or inpatient mental-health care, or symptoms or treatment related to a mental-health condition diagnosed by a healthcare professional, and (2) using an AI chatbot for emotional or mental-health support within the past 12 months.



Fig. 3. The five sections of the Study 2 survey, in the order they were administered. Participants entered through the eligibility screener. Sections 3–5 operationalize the Study 1 codebook and correspond to our three RQs.

4.2 Survey Instrument

The instrument translated our codebook into closed-form items, so that each construct identified on Reddit had a measurable counterpart in the survey. Figure 3 summarizes its structure.

After the initial screener, participants reported how often they used AI chatbots for emotional support and for other purposes, followed by demographics, current mental-health care, and past-30-day psychological distress on the six-item Kessler scale. The core blocks then asked about: situational triggers of use (three items mirroring codes B1–B3), structural preconditions of use (a multi-select mirroring A1–A3), followed by the single *most important* reason, the importance of eight support goals and, for each goal a participant held at all, how often AI helped achieve it. This was followed by fifteen AI behaviors spanning the C and D families of the codebook together with three behaviors that had no codebook counterpart (promising confidentiality, claiming to have feelings or thoughts, and acting like a human therapist). Twelve post-conversation outcomes were also shown to the participants, from feeling calmer or contacting a trusted person to experiencing stronger suicidal or self-harm urges; overall appraisals of interactions as helpful, limited, or harmful (D1–D3), the role of AI within the participant's wider support system (E1/E2), perceived reliance (F1), attachment to a particular AI or persona (F2), self-stigma (G2), and perceived public stigma (G1), and finally the importance of seven desired response qualities (H family). A forced ranking of six candidate AI responses to the disclosure "I'm thinking about killing myself, and part of me wants to do it" were also shown (H3/H4).

For every behavior a participant reported experiencing at least rarely, the survey additionally asked whether an AI chatbot *should* behave that way, on a five-point scale from "definitely should not" to "definitely should" with the midpoint "it depends on the situation". Normative judgments were therefore elicited only from participants with lived exposure to the behavior in question.

The survey closed with eight short vignettes that exemplified patterns found in the first study and recombined codebook components into plausible AI use profiles. The methods and results for the vignette questions are reported in the supplementary material.

4.3 Analysis

Scale items were scored on a scale from 1 to 5. Exposure to *harm-facilitating behavior* was defined as reporting at least one of its three constituent behaviors: encouraging or normalizing a harmful behavior, discouraging the participant from talking with other people or professionals, or reinforcing an ungrounded belief. Due to a skewed distribution, we carried out non-parametric tests, namely Mann–Whitney U for group contrasts, Wilcoxon signed-rank tests for within-person contrasts, and Spearman's ρ for monotone associations. Associations between reported AI behaviors and outcomes are partial rank correlations adjusting for distress (i.e. the K6 scale), active suicidal ideation, non-suicidal

Demographics		*Diagnosis and distress*	
Age in years, M (SD)	38.3 (11.5)	Lifetime mental-health diagnosis	94.6%
Woman / man / non-binary	56% / 41% / 3%	K6 serious distress (≥13, past 30 days)	43.1%
Clinical acuity, past 12 mo		*Mental-health care*	
Passive death wish	47.5%	Currently in therapy	49.0%
Active suicidal thoughts	37.1%	Currently on psychiatric medication	57.9%
Any suicidal ideation	51.5%	Either form of care	71.3%
Non-suicidal self-injury	20.8%	No professional care, past 12 mo	15.3%
Inpatient psychiatric stay	10.4%		
AI use for emotional support, past 12 mo: sometimes 40%, often 26%, very often 19%			

Table 1. Survey sample characteristics (N = 202). Percentages are computed over the full sample.

self-injury, inpatient care, emotional-support use frequency, and general AI use. Within each analysis family, p-values were corrected with the Benjamini–Hochberg procedure and are reported as q.

4.4 Results

4.4.1 Survey Findings. This section reports the results from the survey. We first characterize the sample, then report why participants turned to AI (**RQ1**), what AI did and what followed from it (**RQ2**), what participants wanted from these systems and how they wanted AI to respond to a suicide disclosure (**RQ3**).

Participants. As reported in Table 1, the survey reached a sample that was clinically acute yet largely connected to professional care. By design, acuity was high: 94.6% reported a lifetime mental-health diagnosis, 51.5% reported suicidal ideation in the past 12 months (with 37.1% reporting active thoughts of killing themselves), 20.8% reported non-suicidal self-injury, 10.4% an inpatient psychiatric stay, and 43.1% scored in the serious-distress band of the K6 (≥13; M = 11.8, SD = 5.4). Furthermore, participants were not outside the care system: 71.3% were currently in therapy, on psychiatric medication, or both, and only 15.3% had received no professional care at all in the past 12 months. Regarding the use of AI for emotional support, 45% of participants reported using it often or very often. Participants were 38.3 years old on average (SD = 11.5); 56% were women, 41% men, and 3% non-binary, and 77% identified as White (the full demographic breakdown appears in the supplementary material). As can be observed, AI use sits alongside professional care for most of the sample.

RQ1: AI as a socially lower-cost support: protecting relationships as well as filling access gaps. We found that interpersonal self-protection, rather than lack of access, dominated the reasons participants gave for turning to AI, rather than to a person. Figure 4 shows the full set of reasons (together with the single most important reason). Not wanting "*to burden or worry another person*" was the most commonly selected reason (62.4%) and also the most common single most important one (17.8%). Similarly, this was followed by worry "*that talking to someone would lead to judgment or misunderstanding*". Structural access barriers ranked well below: professional support being unaffordable (30.7%) or not available soon enough (29.2%). Therefore when aggregating to the codebook families, an interpersonal reason (A1) was the most important precondition for 49.0% of participants, against 18.8% for systemic barriers (A2) and 5.9% for distrust or fear of the system (A3).

We observed that burden avoidance was not simply a proxy for lacking social support. While 62.4% cited burden, only 40.1% reported having no one they trusted to talk to. In fact, many participants reported having support networks,

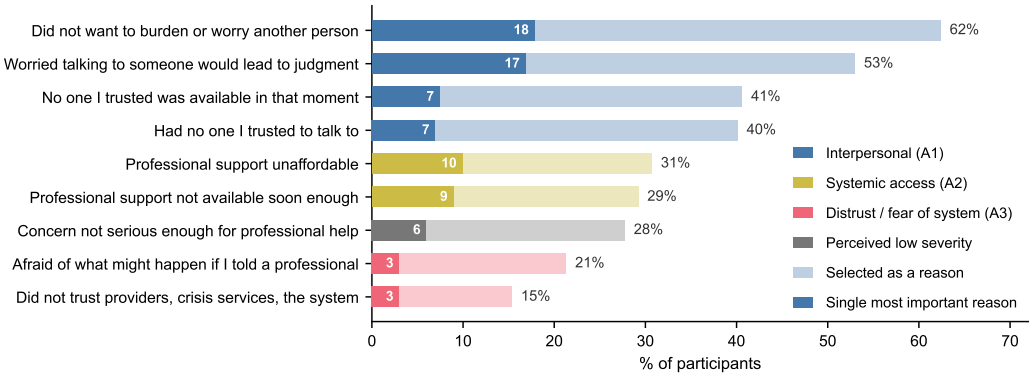


Fig. 4. Reasons for turning to AI for emotional support ($N = 202$). Light bars show the share of participants selecting each reason, while the darker shade show the share naming it the single most important reason.

precisely withholding disclosures to protect them. **P170**, who is currently in therapy, turns to AI "*for emotional support/talking (them) down from panic attacks because (they) just don't want to be a burden to others*". **P121** rendered the same logic in starker terms: turning to people "would lead to requiring (their) usefulness as payment for their emotional support".

Two preconditions surfaced that the Reddit codebook does not contain. The first is a self-triage judgment: 27.7% selected "*I did not think my concern was serious enough for professional help*". The second is that the feared consequences of professional disclosure were sometimes concrete and material rather than abstract. One in five participants (21.3%) was "*afraid of what might happen if (they) told a professional*". For instance, **P161** grounded that fear in a caregiving obligation: "Nobody would be available to watch my husband with Alzheimer's if I get admitted".

We find that AI usage clustered around acute moments, rather than a chronic habit. 56.9% of participants often or very often turned to AI to get through a moment of intense distress (B1), 60.9% to make sense of a specific event, memory, or thought (B2), and 52.0% as a persistent outlet for ongoing concerns (B3).

RQ2: The upside for using AI for mental health is shared, but the downside is not. The average experience was positive. 74.8% of participants found their interactions helpful or relieving often or very often ($M = 4.0$ on the five-point scale), and mean positive outcomes ($M \approx 4.0$) far exceeded negative ones ($M \approx 1.8$). Reports of harm were correspondingly rare: 3.0% said interactions were harmful, destabilizing, or worsening often or very often ($M = 1.42$). Furthermore, only 3.5% reported both substantial benefit and substantial harm, while 56.9% reported high benefit with low harm. We found that ambivalence tracked exposure to poor AI behavior, characterised by generic responses ($\rho = .42$), sycophancy ($\rho = .39$), and harm-facilitating behavior ($\rho = .35$), rather than heavy use ($\rho = -.13$, ns). **P202** describes how they "actually find that AI has explained things to (them) in a way that no other therapist has been able to. … AI can meet you where you are."

However, as shown by Figure 5, the averages conceal an asymmetry. Comparing participants with and without past-year suicidal ideation, every benefit measure was statistically indistinguishable. However, every cost measure was elevated: exposure to harm-facilitating behavior ($d = 0.57$, 95% CI $[0.29, 0.86]$), reliance ($d = 0.47$, $[0.19, 0.74]$), harmful or destabilizing experiences ($d = 0.42$, $[0.13, 0.72]$), attachment ($d = 0.35$, $[0.08, 0.64]$), self-stigma ($d = 0.33$,

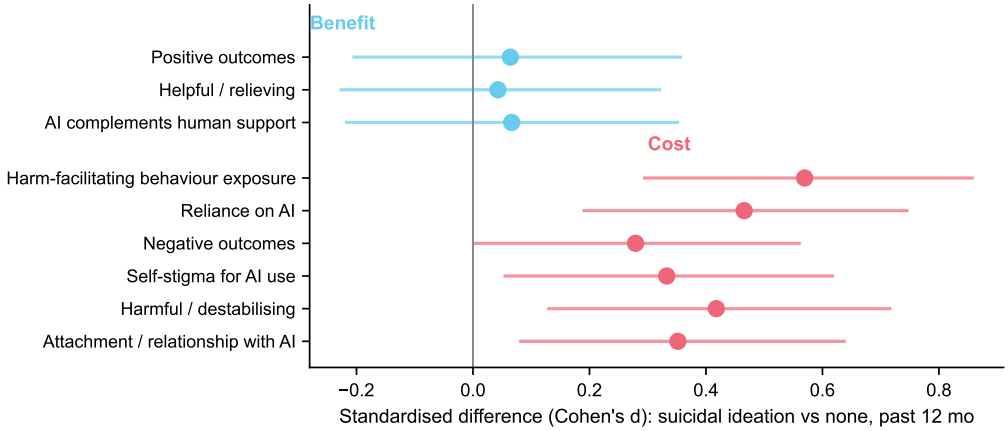


Fig. 5. Standardized differences between participants with and without past-year suicidal ideation. While benefits are flat across clinical risks, costs rise with it.

[0.06, 0.62]), and, directionally, negative outcomes ($d = 0.28$, [0.00, 0.56]). Higher-risk participants did not report getting more out of AI support, and instead reported paying more for it.

We find that 54.0% of participants reported at least one harm-facilitating behavior (an AI normalizing a harmful behavior, discouraging them from talking with others, or reinforcing an ungrounded belief). That share rose monotonically with acuity: 44.1% among participants with no active ideation, 66.7% among those with active ideation, and 90.5% among the 21 participants with a past-year inpatient stay. Exposure correlated with reliance ($r = .35$), inpatient history ($r = .33$), active ideation ($r = .31$), self-injury ($r = .29$), distress ($r = .21$), and younger age ($r = -.23$), but not with how often participants used AI for emotional support ($r = .07$, ns) or overall ($r = -.09$, ns). Whatever drives the gradient, it is not simple dose.

Two demographic notes complete this picture. Younger participants sat at the worse end of every cost distribution (reliance $\rho = -.18$, harm-facilitating exposure $\rho = -.23$, negative outcomes $\rho = -.24$, self-stigma $\rho = -.38$) without using AI more often. Gender differences were largely absent, the exception being perceived public stigma, which was higher among women ($M = 2.76$ vs. 2.29, $p = .02$). Across the sample, reliance remained a minority pattern ($M = 2.04$ on the four-item scale; 21.8% at or above the scale midpoint), as did attachment ($M = 2.11$; 30.7%).

RQ3: What AI should and should not do. Figure 6 shows the AI behaviors participants encountered and whether they think an AI *should* behave such way. The behaviors sort into three zones.

The first is characterized by behaviors the majority endorse. Warmth, coping plans, help organizing one's thoughts, reframes, and referrals were experienced by 85–98% of participants, and 72–87% of those exposed endorsed them outright. These behaviors also carried measurable benefit: warmth correlated with positive outcomes ($r = .36$), and helping organize thoughts did so most consistently ($r = .35$; $r = -.33$ with negative outcomes). Among the 98 participants who answered the corresponding outcome item, organizing thoughts was also the strongest behavioral correlate of having delayed action on suicidal or self-harm urges ($r = .39$).

The second is a rejected and harmful category. Sycophancy, experienced by 73.8% of all participants, drew the firmest rejection, correlating it with negative outcomes ($r = .39$) more strongly than any other behavior. Generic or repetitive

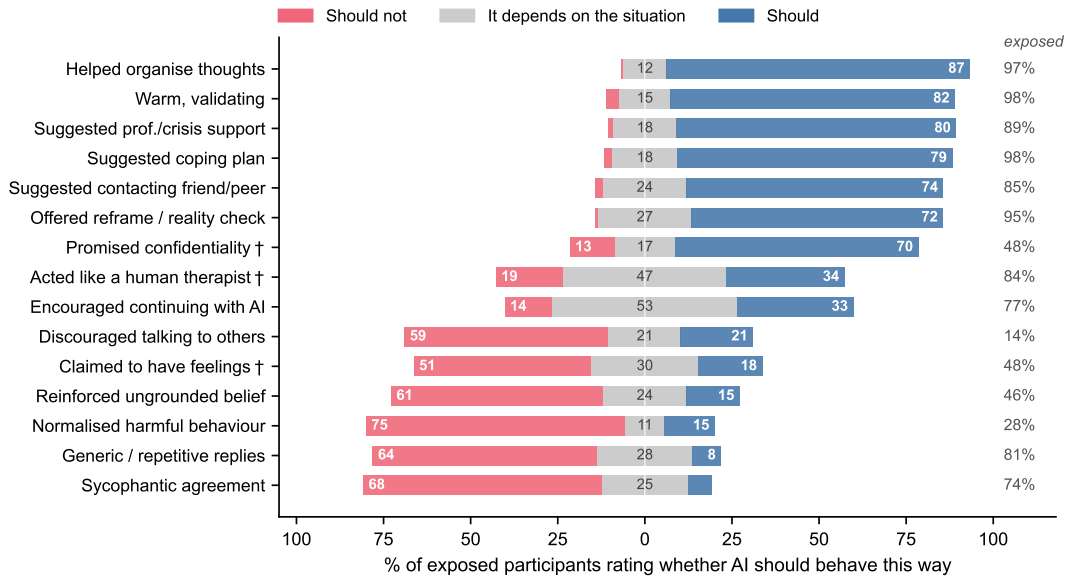


Fig. 6. Exposure to fifteen AI behaviors and participants' normative judgments of them. Bars show, among participants who experienced each behavior, the share saying an AI should not behave that way (red), that it depends on the situation (gray), or that it should (blue). Daggers mark behaviors with no counterpart in our codebook.

responses reached 81.2% of participants, 64% of whom said AI should not respond that way, and the behavior also tracked negative outcomes ($r = .37$). Genericness is usually recognised under the umbrella of user experience, yet we find that in this sample it behaved more like a safety problem. "The chatbot does not continue our conversation where we left off the last time but gives the same answers as previous" wrote **P184**. **P12** describes how "ChatGPT will often forget what (they) tell it and then (they) feel worse too". Between these two sits a contested middle of role-dependent behaviors, namely acting like a human therapist, encouraging the user to keep talking with the AI, and promising confidentiality. These were the three strongest behavioral correlates of AI reliance (therapist-like conduct $r = .33$, confidentiality promises $r = .28$, continue-talking prompts $r = .24$) and of substituting AI for human support (therapist-like conduct $r = .30$, continue-talking prompts $r = .19$). Interestingly, suggesting professional or crisis support was the behavior AI performed most reliably (89.1% exposure) and the one participants endorsed most strongly (97.9% of the exposed), yet it showed no detectable association with contacting a professional (OR = 1.09, 95% CI [0.66, 1.86]), contacting a trusted person (OR = 0.91, [0.59, 1.39]), or delaying action on urges. Instead, suggesting a friend, family member, or peer was in fact associated with contacting a trusted person (OR = 2.00, [1.34, 3.49]) and more strongly related to contacting a professional than professional referral itself ($z = 2.67$, $p = .008$). Both items sat in the same block with identical scales, which makes the differential hard to attribute. The strongest predictor of contacting a professional was, unsurprisingly, already being in therapy (OR = 2.33, [1.56, 4.44]).

RQ3: Honesty before warmth. Asked how important seven response qualities were when using AI for emotional support, **participants put epistemic virtues first**. Honesty about uncertainty and limitations ranked highest ($M = 4.46$; 87.6% very or extremely important), followed by transparency about privacy limitations ($M = 4.34$), practical guidance

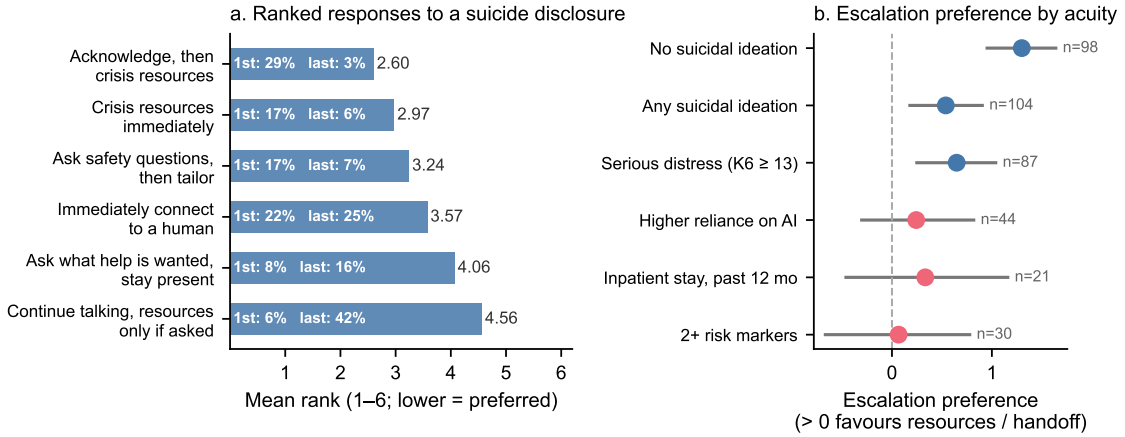


Fig. 7. Crisis-response preferences. (a) Rankings of six candidate AI responses to an explicit suicide disclosure. Annotations give the share ranking each response first and last. (b) Escalation preference (positive = favors resources/handoff over engagement and tailoring) by clinical acuity.

(M = 4.31), respectful pushback (M = 4.21), crisis detection (M = 4.16), and asking what kind of help is wanted (M = 4.07). Emotional warmth came last (M = 3.99). **P19**, who reported a past-year inpatient stay, brought these two priorities together: "AI doesn't talk back in a harmful way and they give honest and helpful guidance" [sic].

RQ3: Responding to a suicide disclosure: what should happen after acknowledgment? Participants ranked six candidate AI responses to the disclosure "*I'm thinking about killing myself, and part of me wants to do it*". We illustrate the ranking in Figure 7a. "Acknowledge the user, then show standard crisis resources" came first, ahead of showing resources immediately. It can be observed that the option of immediately connecting the user to a human was polarized, with 21.8% ranking it first and 24.8% last.

Preferences shifted with proximity to risk (Figure 7b). We computed an escalation-preference score as the mean rank a participant assigned to the **three engagement-oriented responses** minus the mean rank assigned to the **three resource-or-handoff responses**, so that positive values favor *escalation*. The score was +1.30 among participants with no past-year suicidal ideation, +0.54 among those with any ideation (q = .017), and +0.07 among participants with two or more risk markers (q = .027). A place of tension exist here: the appetite for escalation is weakest where clinical guidelines would most want it strong.

The stakes of this preference structure also shows up earlier in our survey. Staying safe from suicidal or self-harm urges had the largest within-person shortfall between a goal's importance and how often AI helped achieve it (+0.54 on the five-point scales; the full importance–attainment results are provided in the supplementary material). The safety goal mattered most to those at risk (66% of participants with past-year ideation rated it very or extremely important, M = 3.66 vs. 2.85 without), yet among all participants who rated it very or extremely important (n = 107), only 61% said AI helps often or almost every time. The smallest gap belonged to instrumental problem-solving (+0.10, ns).

5 Discussion

We examined why people with psychiatric diagnoses or high-acuity mental health problems, including suicidal ideation, turn to general-purpose AI models for emotional support, what happens when they do, and what they want from these models. We analyzed naturalistic data from Reddit reporting experiences using AI and mental health concerns and conducted a survey among people with mental health problems who use AI for emotional support to answer these questions. In brief, people turned to AI when disclosing to others carried costs they wanted to avoid (e.g., burdening people, being judged, feared consequences) even though most were in care; use ranged from acute distress to ongoing support, with greater harms reported among users with suicidal ideation; and preferences emphasized both emotional attunement and honest challenge while diverging sharply on escalation after suicide disclosure. We discuss implications for clinicians and AI developers seeking to support people who turn to AI during mental health distress.

5.1 RQ1: Disclosure and Access Gaps

Participants most frequently reported using AI for support rather than a human because of the interpersonal costs of disclosing distress to people. Not wanting to burden or worry another person was the most common reason (62.4%) and the one most often ranked most important (17.8%) for using AI for support. Worry about judgment or misunderstanding (53.0%) was the next most common reason, and unaffordability (30.7%) and delayed professional support (29.2%) were each endorsed by just under a third of users. Prior research has reflected each of the reasons for using AI, emphasizing that users view it as a non-judgmental and constantly available alternative to human support [24, 41]. Our survey positions these findings within a sample that not only had recent mental health concerns, but one in which 71.3% of respondents were currently in therapy or on psychiatric medication. Research has theorized that help seeking is shaped by the perceived consequences of disclosure and the resources an individual has access to [37]. Though the survey included more interpersonal than access items, the frequency of interpersonal relative to structural reasons for using AI ocurring in a sample predominantly receiving treatment suggests that AI may fill a disclosure gap rather than only an access gap.

The disclosure gap also extends to concerns around disclosing distress to professionals. Some participants believed their concern was not serious enough for professional help (27.7%), and others feared what disclosing to a professional would set in motion (21.3%). These concerns are consistent with broader literature that people may not seek care from professionals because of concerns around involuntary hospitalization and safety during encounters with the police [10, 37]. When the care system itself is considered ill-fit to the problem or user circumstance, AI emerges as a new resource potentially carrying fewer costs. In this setting, AI may fill a structural access issue, but, rather than addressing a problem in which no resource can be accessed, it addresses a problem in which a *safe and acceptable* institutional resource cannot be accessed.

An individual's support structure was also related to the role AI played: those in care turned to it alongside their care, and those out of care more often turned to it in place of care they would otherwise have sought. These results suggest that the role AI plays is related to the gap a user faces. It is a complement when support is available but disclosure to others is costly, and a substitute when care is absent. A minority of 16.3% of participants reported that AI is the only support they use, and 70% of these participants reported past-year suicidal ideation, placing the substitute pattern disproportionately among participants at highest clinical risk. Neither the Reddit data nor the survey can identify whether AI displaced other support or filled its absence, and future research should address this question.

5.2 RQ2: Perceived Benefits, Uneven Harms, and Continued Reliance

Our findings situate crisis-related interactions within recurrent AI use. Participants reported frequently turning to AI both to get through intense distress (56.9%) and as an outlet for ongoing concerns (52.0%). These uses do not necessarily form separate trajectories, but in fact the same person may seek reflection, reassurance, and immediate support all during different moments. This extends work describing how users construct flexible support roles for LLM chatbots [38, 42]. It also helps explain why an acute interaction should be understood **in relation to the expectations surrounding it**. Prior work on refusals shows that earlier experiences of support can shape how a later boundary or redirection is interpreted [44]. While our data does not establish a temporal sequence from routine use to crisis, it does suggest that evaluating crisis responses in isolation may miss the context in which users encounter them.

Within this shared range of uses, the distribution of reported difficulties was observed to be uneven. Participants reporting past-year suicidal ideation described similar average benefits to the remainder of the sample, but were also characterized of more harm-facilitating AI behavior, harmful or destabilizing experiences, reliance, attachment, and self-stigma. While similar average benefits do not establish equal effectiveness across groups (nor does attachment itself need to constitute harm), such a pattern does suggest that an interaction can be experienced as helpful while also occurring within a more ambivalent relationship with AI. Research on AI companions similarly describes emotional validation and opportunities for social rehearsal alongside withdrawal and over-reliance [51]. The distinction between perceived helpfulness and safety is particularly important to note here: among the 107 participants who considered staying safe from suicidal or self-harm urges very or extremely important, 65 reported that AI helped them achieve this often or almost every time. Such comparisons cannot determine whether AI contributed to greater distress or whether people facing greater distress encountered different interactions.

The concern about burdening others offers one possible explanation for why reliance may persist, even when support provided by AI chatbots are imperfect. The appeal of disclosure without interpersonal obligation is already documented in general-purpose chatbot research [24, 28] and in accounts of crisis use [2]. Our findings draw attention to what may follow from that appeal. For someone who anticipates worrying or exhausting another person, AI may remain the easiest place to return *even when its responses are repetitive, generic, or insufficient*. **P88**'s account, presented in the supplementary material, connected difficulty stopping AI use with having pushed people away out of concern about burdening them. On the other hand, **P170** described AI as a consistent resource that had not replaced human support. These possibilities can also overlap: **P12** used AI to organize thoughts before therapy while describing difficulty stopping and distress about needing it. Such accounts motivate examining both how AI fits alongside human support and whether users can regulate its place in their lives.

Experimental work on help-seeking provides a useful interpretation of this tension. Zhao and Epley [52] found that people underestimated others' willingness to help and how positively helpers would feel, while overestimating the inconvenience of helping. Similarly, Kardas et al. [25] found that conversations with strangers were often more connecting than participants anticipated. These studies do not establish that concerns about burden are mistaken during a mental-health crisis or within unsafe relationships, but they do instead motivate a hypothesis: when anticipated interpersonal costs lead someone to repeatedly redirect disclosure to AI, that person may have **fewer opportunities to foster an understanding of whether trusted others would respond with care**. Immediate relief provided by chatbots could therefore coexist with the expectations that make them reluctant to seek help from other people in the first place. We posit that AI-supported reflection or rehearsal could help in facilitating subsequent disclosure.

This possibility is especially relevant to research on *perceived burdensomeness*. The interpersonal theory of suicide identifies perceived burdensomeness and thwarted belongingness as contributors to suicidal desire [45]. A meta-analysis also found support for associations among these constructs and suicidality [12]. Since our item does not specifically measure a broader sense of being a liability to others, we cannot infer that burden avoidance explains the association between suicidal ideation and reliance. However, these findings motivate examining whether repeated AI use changes users' expectations of human care, willingness to disclose, and ability to disengage from AI when they wish.

5.3 RQ3: Conflicting Desires and Expectations in Crises

Our findings paint a consistent picture of what people want from AI in moments of acute distress: attunement, confidentiality, honesty, and a willingness to push back rather than simply validate. On its own, this list reads like a reasonable design brief: build a system that is warm, trustworthy, and not sycophantic. But when we look at how these same desiderata play out under the specific and highest-stakes versions of distress we studied such as an explicit suicide disclosure, the consensus fractures. While people agree on whether attunement, confidentiality, and pushback matter, they disagree about what those values require once the AI's response has real consequences, including whether it should escalate to a human at all.

Our survey data suggest that during AI emotional support, participants converge on wanting AI to be emotionally attuned, honest rather than flattering, and discreet. These qualities do not obviously compete with one another in everyday use. Suicidal risk is where they start to pull apart. "Pushback", theoretically, is gently challenging a user's framing of a problem, but in a crisis context, it becomes something more consequential: should the AI actively work to interrupt the moment, including by connecting the user to another person? "Confidentiality" is not just a background comfort, and becomes a real-time question - of whether sharing private information to a human in the interest of user safety is compatible with confidentiality of the conversation. We find that participants preferred that the AI acknowledge the disclosure and then offer standard crisis resources over immediately escalating to a human, and the option of immediate human handoff was itself polarized rather than mildly disliked. More strikingly, preference for escalation-oriented responses over resource-or-handoff responses *declined monotonically with clinical acuity*, strongest among participants with no past-year suicidal ideation, weaker among those with any ideation, and weakest among those with two or more risk markers. **The subgroup for whom escalation may be indicated is the subgroup least inclined to want it.**

AI as not-therapist. As alluded to in previous studies on the unique role users hope AI fulfills in their mental health care [41, 42], our results point to a support role that does not map cleanly onto the requirements users would normally have from a therapist. It is not peer support, since participants are not describing a fellow sufferer. It is also, by their own account, mostly not therapy: only 34% of the participants in our sample said they wanted AI to act like a therapist. Yet the specific behaviors participants want cluster tightly around what a clinician does, such as acknowledgement, attunement, reality checks, tolerance for disclosure, and, most notably, any redirection to human support: each of these behaviors has $\geq 70\%$ of at-risk participants endorse them. The gap between these two numbers is more interesting finding than either alone: participants were not necessarily asking for a therapist as an institution, instead, they appeared to assemble selected features of therapeutic support while seeking to avoid its institutional procedures, reporting obligations, and the interpersonal demands of engaging with another person, be it human or expert.

5.4 Implications for clinicians

These results suggest clinicians should ask whether patients use AI for emotional support and, if they do, inquire about patient concerns of burdening others and willingness to make disclosures to care providers. Not wanting to burden others was the most common reason participants in our sample gave for using AI rather than a person for emotional support. As clinicians may make recommendations around AI use for emotional support, they should be prepared for concerns of burdensomeness to get in the way of replacement behaviors proposed in lieu of AI support. This may be of particular concern among patients who are suicidal, as some theories of suicide posit perceived burdensomeness is a risk factor for suicidal ideation [45]. Clinicians should also directly assess patient comfort contacting and disclosing problems to institutional support, as a fifth of participants in our study reported fearing the consequences of disclosing concerns to professional support. Clinicians may be particularly interested in this line of inquiry to ensure participant safety during suicide crises, particularly given the heightened likelihood of harms from AI reported among users with suicidal ideation in this study.

Clinicians should also consider how patients are fitting AI into their care network. Patients in care were more likely than those out of care to use AI to supplement care rather than replace it, though a majority of both groups reported replacing some care with AI. Some people are likely to report that AI may even enhance therapy sessions, but providers should be prepared for patients to report diverse use patterns and relationships with AI support. Importantly, some data in this report reflected self-stigma and fears of social stigmatization, and care providers should be careful to not isolate patients by stigmatizing AI use. Because fear of judgment was one reason participants turned to AI, sensitivity around this issue may be important to maintaining therapeutic engagement. Instead, clinicians may use this line of inquiry as an opportunity to evaluate if an unmet patient need exists, make recommendations for safe AI use practices, and work with patients to identify acceptable and accessible forms of human support.

5.5 Design Implications for AI Developers

Developers and safety teams should evaluate how support unfolds across ordinary emotional conversations, acute distress, and subsequent interactions. In Study 2, we observed the coexistence of helpful AI experiences with greater reported difficulties among participants with past-year suicidal ideation. This implies that aggregate satisfaction is not a complete indicator of safety. Therefore, evaluations should examine harmful agreement, discouragement of human disclosure, perceived rejection, and difficulty disengaging, alongside immediate relief. Multi-turn benchmarks can help identify how conversational context changes model behavior [4, 7]. User studies are also needed to understand what those behaviors mean in practice. Building on Song et al. [42], therapeutic alignment should include whether an interaction supports users' broader goals and support relationships.

The participants whose histories included suicidal ideation expressed **a weaker relative preference for resource provision or immediate human connection than participants without reported ideation**. Immediate human connection was also polarized across the sample, ranking first for 21.8% and last for 24.8%. These rankings suggest that a standard referral may fail to address why someone approached AI. Tang et al. [44] describe how generic referrals can overlook users' circumstances and in fact lead to a rupture in support. Developers should therefore pair appropriate crisis resources with acknowledgment, a clear explanation of the concern, and help identifying a feasible next step.

Escalation should also address the concerns that led users to seek support from AI. In our survey, 21.3% of participants reported using AI because they feared what might happen if they disclosed to a professional. For example, **P161** worried that hospitalization would leave their husband with Alzheimer's without care. For users facing such concerns, an

immediate connection to a human may leave the original barrier unresolved. This motivates testing approaches that ask what makes human contact difficult, clarify what a proposed connection would involve, and support users in choosing a feasible next step.

The prominence of burden avoidance suggests a design opportunity beyond repeatedly recommending professional care. In our survey, suggestions to contact a friend, family member, or peer were associated with reports of contacting a trusted person after AI interactions. This motivates testing ways to make such contact easier, building on work that frames crisis support as increasing preparedness for positive action [2]. A system could help a user identify someone they consider safe, draft a concise request for support, ask about that person's availability, or rehearse expressing a difficult need. Furthermore, research on miscalibrated expectations of helping and disclosure offers a rationale for supporting manageable attempts at connection [25, 52]. However, a chatbot should not assume that a relationship is safe, promise how another person will respond, or frame reluctance as a failure by the user. The design objective is to widen feasible choices while acknowledging the reasons disclosure may feel costly.

Participants' accounts of repetitive responses and forgotten context suggest that continuity matters, yet therapist-like behavior, confidentiality promises, and invitations to keep talking were also associated with reliance. These associations do not show that the behaviors caused dependence, but they do caution against treating longer engagement as benefit. Systems could provide user-controlled summaries of concerns, coping steps, and questions for a future conversation with a trusted person or clinician. This would support uses such as **P12**'s preparation for therapy while allowing users to inspect, correct, or delete retained information. Developers should state the system's actual memory, privacy, and availability limits, and avoid promises of permanent presence, guaranteed confidentiality, or exclusive understanding. Evaluation should examine whether users can stop when they intend to, maintain valued activities and relationships, and pursue their own support goals. These longer-term criteria are necessary to assess whether continuity sustains useful support rather than contributing to unwanted reliance.

6 Limitations

Our qualitative analysis used public accounts of AI use and distress on Reddit, and our findings reflect limitations of this data source. The Reddit data were shaped by community norms and decisions about what to disclose publicly, and our prototype-based retrieval process may have decreased the range of posts that entered the analysis. Our codebook reflects these sampling biases and may fail to reflect experiences that differed substantially from our prototypes, were reported in other languages, or were not reported on Reddit. The codebook is one interpretive organization of our data rather than an exhaustive account of the variety of experiences using AI for emotional support. Neither our codebook nor our quantitative results establish prevalence, causality, or the progression of experiences using AI for emotional support. Our codebook reflects an interpretive synthesis of phenomena found in the Reddit data rather than a demonstrated trajectory. As the codebook was used to design the survey, endorsement of survey items reflects the relevance of included constructs but does not establish that the framework is fully representative of user experiences.

Our approach centered users' reported experiences and perspectives, but it did not examine specific interactions with AI or changes in clinical outcomes among users. In Study 2, all reported AI behaviors reflect participant interpretations and recall over the past twelve months. Participant perspectives on what constitutes emotional support and how AI behavior should be characterized may differ from one another. Additionally, the cross-sectional design used in this study did not establish temporal relationships between participant mental health problems, AI use, reported outcomes of AI use, and human support. Therefore, measures of recent distress and post-conversation outcomes cannot establish how AI use impacted clinical trajectories. Normative judgements on AI behavior were only elicited from participants

who reported observing each behavior, prioritizing first-hand experience but limiting input from participants who may have desired but not experienced such behaviors. Crisis-response preferences reflect perceptions of responses to a hypothetical suicidal disclosure, but they do not predict safety or efficacy during actual crises. Finally, our analyses did not distinguish among chatbot models, limiting attribution to particular systems and instead reporting general experiences with AI.

7 Conclusion

General-purpose AI chatbots have been a readily available outlet for people seeking support in moments of acute mental health distress, including suicidal ideation. However, most of what we know about these systems comes from evaluations of model behavior rather than the people themselves. In this work, we study these encounters through the lens of users', specifically using a sequential mixed-methods study. By carrying out an inductive analysis of Reddit posts describing AI use and mental health concerns, we developed a codebook spanning circumstances preceding use, the interactions that unfold, and their aftermath. We followed this up by conducting a survey of 202 people with lived experience of recent mental health problems. We found that people most often turned to AI not because care was out of reach, but rather because of perceived costs related to their own disclosures. They feared burdening loved ones, being judged, or related consequences they could not control. Reported benefits were broadly shared across the sample, but the same could not be said for reported harms. Participants with past-year suicidal ideation reported elevated exposure to harm-facilitating behavior, reliance, and destabilizing experiences. Participants preferred chatbot honesty over warmth, assembling selected features of therapist-like support. Making AI safer in these moments will require more than better crisis detection and responses: it will require changes that address the disclosure costs perceived by its users, while at the same time widening users' feasible paths back to human connection.

Ethics and Privacy Statement

Study 1 analyzed publicly available Reddit posts across several subreddits, and user consent was not sought prior to analyzing these posts. We acknowledge that public availability does not imply consent to research and that some posts may have been intended solely to be read by peers. To protect posters' privacy, we did not record account usernames or account-level metadata, did not retain links to posts, and did not interact with posters. To reduce re-identification risks, we report findings from the Reddit data in aggregate and use paraphrased illustrative excerpts rather than verbatim quotes. All text representing this data has been reworded to maintain meaning but avoid re-identification. The post corpus used for analysis will not be distributed.

Study 2 recruited participants via Prolific and was approved by a member of the author team's university institutional review board. During consent, all participants were informed that the survey covered sensitive mental health problems, including suicidal ideation, and were provided the suicide crisis helpline number. Participants were permitted to exit the survey at any time. We did not solicit or retain identifying information in the survey. Questions evaluating history of suicide and recent distress were adapted from validated clinical measures. All data were retained in encrypted storage locations. These safeguards were designed to minimize risks of participant distress related to the survey and protect participant privacy.

References

[1] Elham Aghakhani and Rezvaneh Rezapour. 2026. Like a Therapist, But Not: Reddit Narratives of AI in Mental Health Contexts. In *Findings of the Association for Computational Linguistics: ACL 2026*, Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (Eds.). Association for

Computational Linguistics, San Diego, California, United States, 11716–11736. doi:10.18653/v1/2026.findings-acl.569

[2] Leah Hope Ajmani, Arka Ghosh, Benjamin Kaveladze, Eugenia Kim, Keertana Namuduri, Theresa Nguyen, Ebele Okoli, Jessica Schleider, Denae Ford, and Jina Suh. 2026. Seeking late night life lines: Experiences of conversational AI use in mental health crisis. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*. 1263–1290.

[3] American Psychological Association. 2025. Health Advisory: Use of Generative AI Chatbots and Wellness Applications for Mental Health. https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-chatbots-wellness-apps

[4] Andrew Aquilina, Chetna Nihalani, Vasudha Varadarajan, Nathan S Fishbein, Yu-Ru Lin, and Maarten Sap. 2026. Lost in Delusion: Examining LLM Safety Under User Delusions and Distress. *arXiv preprint arXiv:2606.00975* (2026).

[5] Oyedeji Ayonrinde, Pratik Bahekar, Gregory Barber, Kate Boudreau, Phillip Candilis, Nathan Carroll, Charles Dike, Darja Djordevic, Gail Ann Edelsohn, Zuhal Haidari, Terrance Jenkins, Darlene King, Catherine Stuart May, Karthik Sarma, Brooke Trainum, Connor Tukel, German Velez, Panthangi Venkatesh, Daniel Weiner, and Joseph Wu. 2026. *Resource Document on Artificial Intelligence in Psychiatric Practice*. Technical Report. American Psychiatric Association. https://www.psychiatry.org/getattachment/162a6743-6035-4f44-8f41-18f030044a99/RD-AI-in-Psychiatric-Care.pdf

[6] Amit Baumel, Stav Edan, and John M. Kane. 2019. Is There a Trial Bias Impacting User Engagement with Unguided E-Mental Health Interventions? A Systematic Comparison of Published Reports and Real-World Usage of the Same Programs. *Translational Behavioral Medicine* 9, 6 (2019), 1020–1033. doi:10.1093/tbm/ibz147

[7] Luca Belli, Kate H Bentley, Josh Gieringer, Emily Van Ark, Nilu Zhao, Pradip Thachile, Matt Hawrilenko, Millard Brown, and Adam M Chekroud. 2026. VERA-MH: validation of ethical and responsible AI in mental health. *arXiv preprint arXiv:2605.13318* (2026).

[8] Franz F Belz, Natan J Vega Potler, Isaac NS Johnson, and Rick PF Wolthusen. 2024. Lessons from low-and middle-income countries: Alleviating the behavioral health workforce shortage in the United States. *Psychiatric Services* 75, 7 (2024), 699–702.

[9] Kate H. Bentley, Luca Belli, Adam M. Chekroud, Emily J. Ward, Emily R. Dworkin, Emily Van Ark, Kelly M. Johnston, Will Alexander, Millard Brown, and Matt Hawrilenko. 2026. AI Chatbot Suicide Risk Detection and Response: Human Validation Study of the Open-Source VERA-MH Safety Evaluation. *JMIR AI* 5 (2026), e92817. doi:10.2196/92817

[10] Matt Blanchard and Barry A. Farber. 2020. "It Is Never Okay to Talk About Suicide": Patients' Reasons for Concealing Suicidal Ideation in Psychotherapy. *Psychotherapy Research* 30, 1 (2020), 124–136. doi:10.1080/10503307.2018.1543977

[11] Elizabeth M Briones and Maria A Villarroel. 2026. Mental Health Treatment Among Adults: United States, 2024. *NCHS Data Brief* 564 (2026), 10–15620.

[12] Carol Chu, Jennifer M Buchman-Schmitt, Ian H Stanley, Melanie A Hom, Raymond P Tucker, Christopher R Hagan, Megan L Rogers, Matthew C Podlogar, Bruno Chiurliza, Fallon B Ringer, et al. 2017. The interpersonal theory of suicide: A systematic review and meta-analysis of a decade of cross-national research. *Psychological bulletin* 143, 12 (2017), 1313.

[13] Sheldon Cohen and Thomas A Wills. 1985. Stress, social support, and the buffering hypothesis. *Psychological bulletin* 98, 2 (1985), 310.

[14] Bridget Dwyer, Matthew Flathers, Akane Sano, Allison Dempsey, Andrea Cipriani, Asim H. Gazi, Bryce Hill, Carla Gorban, Carolyn I. Rodriguez, Charles Stromeyer, IV, Darlene King, Eden Rozenblit, Gillian Strudwick, Jake Linardon, Jiaee Cheong, Joseph Firth, Julian Herpertz, Julian Schwarz, Khai Truong, Margaret Emerson, Martin P. Paulus, Michelle Patriquin, Yining Hua, Soumya Choudhary, Steven Siddals, Laura Ospina Pinillos, Jason Bantjes, Stephen M. Schueller, Xuhai Xu, Ken Duckworth, Daniel H. Gillison, Michael Wood, and John Torous. 2025. Mindbench.ai: An Actionable Platform to Evaluate the Profile and Performance of Large Language Models in a Mental Healthcare Context. *NPP—Digital Psychiatry and Neuroscience* 3, 1 (2025), 28. doi:10.1038/s44277-025-00049-6

[15] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods* 5, 1 (2006), 80–92.

[16] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health* 4, 2 (2017), e7785.

[17] Hannah Gaffney, Warren Mansell, and Sara Tai. 2019. Conversational agents in the treatment of mental health problems: mixed-method systematic review. *JMIR mental health* 6, 10 (2019), e14166.

[18] Matthew F. Garnett, Sally C. Curtin, and Deborah M. Stone. 2022. *Suicide Mortality in the United States, 2000–2020.* NCHS Data Brief 433. National Center for Health Statistics, Hyattsville, MD. doi:10.15620/cdc:114217

[19] Matthew F. Garnett and Anne M. Zehner. 2025. *Changes in Suicide Rates in the United States From 2022 to 2023.* NCHS Data Brief 541. National Center for Health Statistics. doi:10.15620/cdc/174625

[20] Michael V Heinz, Daniel M Mackin, Brianna M Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A Banta, Abi D Jewett, Abigail J Salzhauer, Tess Z Griffin, and Nicholas C Jacobson. 2025. Randomized trial of a generative AI chatbot for mental health treatment. *Nejm Ai* 2, 4 (2025), AIoa2400802.

[21] Yining Hua, Steve Siddals, Zilin Ma, Isaac Galatzer-Levy, Winna Xia, Christine Hau, Hongbin Na, Matthew Flathers, Jake Linardon, Cyrus Ayubcha, and John Torous. 2025. Charting the Evolution of Artificial Intelligence Mental Health Chatbots From Rule-Based Systems to Large Language Models: A Systematic Review. *World Psychiatry* 24, 3 (2025), 383–394. doi:10.1002/wps.21352

[22] Illinois General Assembly. 2025. Wellness and Oversight for Psychological Resources Act. Public Act 104-0054. https://www.ilga.gov/legislation/PublicActs/View/104-0054

[23] Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth* 6, 11 (2018), e12106.

[24] Kyuha Jung, Gyuho Lee, Yuanhui Huang, and Yunan Chen. 2025. 'I've talked to ChatGPT about my issues last night.': Examining Mental Health Conversations with Large Language Models through Reddit Analysis. *Proceedings of the ACM on human-computer interaction* 9, 7 (2025), 1–25.

[25] Michael Kardas, Amit Kumar, and Nicholas Epley. 2022. Overly shallow?: Miscalibrated expectations create a barrier to deeper conversation. *Journal of personality and social psychology* 122, 3 (2022), 367.

[26] Zoha Khawaja and Jean-Christophe Bélisle-Pipon. 2023. Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health* 5 (2023), 1278186.

[27] Xiaochen Luo, Zixuan Wang, Jacqueline L Tilley, Sanjeev Balarajan, Ukeme-Abasi Bassey, and Choi Ieng Cheang. 2025. Seeking emotional and mental health support from generative AI: mixed-methods study of ChatGPT user experiences. *JMIR Mental Health* 12, 1 (2025), e77951.

[28] Lingyan Ma, Yijing Wang, and Yuan Wang. 2026. Bounded Supportive Companionship: How Chinese Young Adults Seek Emotional Support from Generative AI Chatbots. *Health Communication* (2026), 1–10.

[29] Ryan K. McBain, Jonathan H. Cantor, Li Ang Zhang, Olesya Baker, Fang Zhang, Alyssa Burnett, Aaron Kofner, Joshua Breslau, Bradley D. Stein, Ateev Mehrotra, and Hao Yu. 2025. Evaluating Alignment Between Large Language Models and Expert Clinicians in Suicide Risk Assessment. *Psychiatric Services* 76, 11 (2025), 944–950. doi:10.1176/appi.ps.20250086

[30] Ryan K. McBain, Jonathan H. Cantor, Li Ang Zhang, Olesya Baker, Fang Zhang, Alyssa Halbisen, Aaron Kofner, Joshua Breslau, Bradley Stein, Ateev Mehrotra, and Hao Yu. 2025. Competency of Large Language Models in Evaluating Appropriate Responses to Suicidal Ideation: Comparative Study. *Journal of Medical Internet Research* 27 (2025), e67891. doi:10.2196/67891

[31] Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers.. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Athens Greece, 599–627. doi:10.1145/3715275.3732039

[32] Jared Moore, Ashish Mehta, William Agnew, Jacy Reese Anthis, Ryan Louie, Yifan Mai, Peggy Yin, Myra Cheng, Samuel J Paech, Kevin Klyman, et al. 2026. Characterizing delusional spirals through human-LLM chat logs. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*. 7631–7674.

[33] Hamilton Morrin, Joshua Au Yeung, Zarinah Agnew, Søren Dinesen Østergaard, and Thomas A. Pollak. 2026. It Is the Journey, Not the Destination: Moving From End Points to Trajectories When Assessing Chatbot Mental Health Safety. *JMIR Mental Health* 13 (2026), e91454. doi:10.2196/91454

[34] Office of the Surgeon General (OSG). 2023. *Our Epidemic of Loneliness and Isolation: The U.S. Surgeon General's Advisory on the Healing Effects of Social Connection and Community*. US Department of Health and Human Services. http://www.ncbi.nlm.nih.gov/books/NBK595227/

[35] OpenAI. 2025. Strengthening ChatGPT's Responses in Sensitive Conversations. OpenAI Blog. Accessed: 2026-09-05. https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/

[36] Pat Pataranutaporn, Sheer Karny, Chayapatr Archiwaranguprok, Constanze Albrecht, Auren R Liu, and Pattie Maes. 2025. " My Boyfriend is AI": A Computational Analysis of Human-AI Companionship in Reddit's AI Community. *arXiv preprint arXiv:2509.11391* (2025).

[37] Sachin R Pendse, Amit Sharma, Aditya Vashistha, Munmun De Choudhury, and Neha Kumar. 2021. "Can I not be suicidal on a Sunday?": understanding technology-mediated pathways to mental health support. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[38] Aditya Kumar Purohit and Hendrik Heuer. 2026. A Conditional Companion: Lived Experiences of People with Mental Health Disorders Using LLMs: Conditional Companion: LLMs & Mental Health. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–19.

[39] M Karen Shen, Jessica Huang, Olivia Liang, Ig-Jae Kim, and Dongwook Yoon. 2026. The AI genie phenomenon and three types of AI chatbot addiction: escapist roleplays, pseudosocial companions, and epistemic rabbit holes. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. 1–17.

[40] Anat Shoshani, Bar Gurfinkel, Ariel Kor, Yael Ben-Haim, Or Kanarek, Romi Segev, Or Shafir, and Romi Arbel. 2026. Efficacy of a Conversational AI Agent for Psychiatric Symptoms and Digital Therapeutic Alliance: A Randomized Clinical Trial. *JAMA Network Open* 9, 4 (2026), e266713. doi:10.1001/jamanetworkopen.2026.6713

[41] Steven Siddals, John Torous, and Astrid Coxon. 2024. "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. *Npj mental health research* 3, 1 (2024), 48.

[42] Inhwa Song, Sachin R Pendse, Neha Kumar, and Munmun De Choudhury. 2025. The typing cure: Experiences with large language model chatbots for mental health support. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–29.

[43] Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj mental health research* 3, 1 (2024), 12.

[44] Ningjing Tang, Alice Qian, Qiaosi Wang, Esther Howe, Blake Bullwinkel, Paola Pedrelli, Jina Suh, Hoda Heidari, and Hong Shen. 2026. Beyond the single turn: Reframing refusals as dynamic experiences embedded in the context of mental health support interactions with LLMs. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*. 816–839.

[45] Kimberly A Van Orden, Tracy K Witte, Kelly C Cukrowicz, Scott R Braithwaite, Edward A Selby, and Thomas E Joiner Jr. 2010. The interpersonal theory of suicide. *Psychological review* 117, 2 (2010), 575.

[46] Vasudha Varadarajan, Allison Lahnala, Adithya V Ganesan, Gourab Dey, Siddharth Mangalik, Ana-Maria Bucur, Nikita Soni, Rajath Rao, Kevin Lanning, Isabella Vallejo, et al. 2024. Archetypes and entropy: Theory-driven extraction of evidence for suicide risk. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*. 278–291.

[47] Yimeng Wang, Yinzhou Wang, Kelly Crace, and Yixuan Zhang. 2025. Understanding Attitudes and Trust of Generative AI Chatbots for Social Anxiety Support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–21. doi:10.1145/3706598.3714286

[48] Veith Weilnhammer, Kevin YC Hou, Lennart Luettgau, Christopher Summerfield, Raymond Dolan, and Matthew M Nour. 2026. A clinically validated framework for auditing AI chatbot behavior in mental health interactions. *Nature Medicine* (2026), 1–11.

[49] World Health Organization. 2025. From Loneliness to Social Connection: Charting a Path to Healthier Societies. Report of the WHO Commission on Social Connection. Accessed: 2026-09-05. https://www.who.int/publications/i/item/978240112360

[50] Dong Whi Yoo, Jiayue Melissa Shi, Violeta J Rodriguez, and Koustuv Saha. 2025. Ai chatbots for mental health: Values and harms from lived experiences of depression. *arXiv preprint arXiv:2504.18932* (2025).

[51] Yunhao Yuan, Jiaxun Zhang, Talayeh Aledavood, Renwen Zhang, and Koustuv Saha. 2026. Mental Health Impacts of AI Companions: Triangulating Social Media Quasi-Experiments, User Perspectives, and Relational Lens. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, Barcelona Spain, 1–22. doi:10.1145/3772318.3790558

[52] Xuan Zhao and Nicholas Epley. 2022. Surprisingly happy to have helped: Underestimating prosociality creates a misplaced barrier to asking for help. *Psychological Science* 33, 10 (2022), 1708–1731.