International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(21s) 2026 pp. 1917-1929
https://doi.org/10.70917/ijcisim-2026-5968
Article

AI Journaling and Companion Apps for Mental Health: Mechanisms, Outcomes, and Safe Design Principles

Sridhar Rao Muthineni

Optum Services Inc., USA
ORCID: 0009-0002-5382-0435

Abstract: AI-powered journaling and companion apps reach millions daily, yet most lack regulatory oversight and long-term safety data. This paper synthesizes efficacy and harm evidence from randomized controlled trials (RCTs) and observational studies, maps mechanistic pathways, and proposes a five-principle safe design framework, covering reality anchoring, use moderation, vulnerability detection, privacy by design, and clinician integration. Short-term benefits are real but modest; harm concentrates in heavy users and vulnerable subgroups. Short-term benefits are documented across both categories, though empirical evidence for AI journaling applications remains limited relative to companion app research, and long-term outcomes for both remain insufficiently characterized. Regulatory implications for the FDA digital health framework and the EU AI Act are discussed.
Keywords: AI Journaling, AI Companion Apps, Mental Health, Emotional Well-being, Safe Design, Human-AI Interaction

1. INTRODUCTION

The mental health landscape is undergoing a structural transformation driven by artificial intelligence (AI). Two consumer-facing categories have emerged as particularly consequential: AI journaling applications, which offer structured self-reflection, emotional pattern detection, and AI-generated insight; and AI companion applications, which provide persistent, non-judgmental conversational relationships at scale. As of early 2026, more than 40 applications are listed under 'AI journal' on the iOS App Store and on companion platforms such as Replika and Character. AI collectively reports tens of millions of active users [1]. In many countries, these applications now reach more individuals experiencing mental health concerns than the formal mental health system does. The clinical evidence base, however, has not kept pace with deployment. The central problem is the absence of a systematic, evidence-grounded framework that accounts for both the therapeutic potential and the harm pathways of these technologies in a form that developers, clinicians, and regulators can operationalize. A small number of high-quality randomized controlled trials (RCTs) exist, most notably the Flourish multi-institutional study [6] and the Fang et al. longitudinal RCT [5], but they are short-term, conducted on convenience samples, and generally unable to detect subgroup harms with adequate statistical power. Meanwhile, adverse signals are accumulating. Laestadius et al. [7] documented the emotional dependence and psychological harm of intensive Replika users. Additionally, the literature describes cases of AI chatbot delusional reinforcement [4]. A CHI 2026 study triangulating two years of social media data found that heavy AI companion use coincided with increased linguistic markers of distress [2]. These signals are not sufficient to condemn the entire category, but they are sufficient to demand more rigorous analysis.

This paper makes three contributions. First, it systematically decomposes the mechanistic pathways through which AI journaling and companion apps affect mental health, distinguishing expressive-processing effects, AI-augmentation effects, social-presence effects, and behavioral-activation effects. Second, it synthesizes the current efficacy and harm evidence, including dose-response heterogeneity and subgroup risk profiles. Third, it proposes an


Copyright: © 2026 by the authors

operationalizable safe design framework grounded in this evidence, with implications for developers, clinicians, platform operators, and regulators [29].

The analysis draws on peer-reviewed sources through mid-2026. The contributions of this paper are threefold: (1) a structured taxonomy of mechanistic pathways through which AI mental health apps affect outcomes, explicitly distinguishing journaling and companion app mechanisms and their differential risk profiles; (2) a synthesis of dose-response heterogeneity and subgroup risk profiles drawn from 23 empirical studies; (3) a six-step, risk-stratified safe design decision framework operationalizable with currently available clinical screening tools including PHQ-9, GAD-7, and RRS, and real-time NLP monitoring technologies; and (4) a formalized risk stratification model that classifies users into low-risk and high-risk tiers based on PHQ-9, GAD-7, and RRS threshold scores, and maps each tier to a prescribed design pathway and monitoring protocol, operationalizable without clinical infrastructure [30].

2. Background and Related Work

A. Expressive Writing as Therapeutic Foundation

The strongest empirical foundation for any journaling-based intervention comes from Pennebaker's expressive writing paradigm, developed over four decades of controlled experimentation [9]. The canonical protocol, 15 to 20 minutes of uninhibited writing about emotionally significant experiences over three to four consecutive sessions, consistently produces measurable reductions in anxiety, improvements in immune markers, and reductions in health clinic visits across more than 100 studies, with a mean effect size of approximately $d = 0.16$. Critically, the mechanism is understood to operate through a combination of cognitive restructuring, narrative meaning-making, and the reduction of active inhibition associated with undisclosed trauma [9]. This literature establishes both the reference condition and the benchmark against which AI augmentation must justify additional benefit. The expressive-writing effect is modest but replicable; any claim that AI-enhanced journaling outperforms simple expressive writing must therefore clear a meaningful empirical bar. It also establishes a key risk condition: rumination, the maladaptive repetitive focus on distress symptoms, is known to attenuate or reverse the benefits of expressive writing in high-trait-rumination individuals. Whether AI prompting and pattern detection amplify or attenuate rumination in vulnerable users is a critical unresolved question that the current evidence base does not address.

B. AI Journaling Apps: Evidence and Gaps

Systematic reviews of mobile mental health applications identify emotion regulation as one of the most consistently targeted outcome domains, with significant small effects for well-being, stress, depression, and anxiety, and a significant medium effect for emotion regulation across 39 RCTs [15]. Most of the literature on journal apps includes studies of general apps that do not have generative AI capabilities. In contrast, few studies have investigated AI improved formats such as prompted or personalized journaling or recognizing patterns between journal entries and generating synthesized insights. A systematic review in JMIR Mental Health in 2024 [16] found promising but inconclusive support for empathic conversational agent designs, but studies were mostly short (6-8 weeks) and used self-report measures.

The three structural gaps most limiting to the field are, first, study duration: most trials run for one to six weeks, while real users engage for months or years. Second, app generalizability: published trials frequently use sponsor-built research applications that may not reflect commercial app behavior. Third, outcome standardization: studies measure loneliness, well-being, and emotional clarity inconsistently, blocking meta-analytic synthesis. To fill these gaps, we need long-term RCTs with standardized outcome batteries that include validated tools like the Difficulties in Emotion Regulation Scale-18 (DERS-18), Patient Health Questionnaire-9 (PHQ-9), and the Ruminative Response Scale (RRS), as well as pre-registered subgroup analyses focused on rumination-prone individuals.

C. AI Companion Apps: Evidence and Gaps

The evidence base for AI companion applications is both more extensive and more contested than that for AI journaling. De Freitas et al. found that across six studies with a total of 600 participants, AI companions reduced

loneliness to the same degree as a conversation with another person [3]. This finding was taken to mean that AI companions have a therapeutic effect. However, a four-week RCT by Fang et al., the most methodologically rigorous trial to date, with 981 participants and voice-based interaction, produced a more nuanced result: while voice chatbots modestly reduced loneliness on average, heavy daily use correlated with greater loneliness, increased emotional dependence, and reduced real-world socializing [5]. This dose-response non-monotonicity is theoretically significant and practically important.

Longitudinal signals add further complexity. Using data from two years of Reddit discussions, interviews, and relational theory, Yuan et al. concluded in a CHI conference paper published in 2026 that Replika is used daily for emotion regulation and linked it to the connection between Replika usage and the use of linguistic indicators of distress. Laestadius et al. found from grounded theory research with Replika users that practicing role-taking and emotional bonding in the chatbot-as-friend relationship, including in the face of negative experiences, is similar to the harms and emotional attachments possible in human relationships. Dohnany et al. use the term 'technological folie à deux' to describe a shared delusion between the human user and the AI chatbot based on mutual sycophancy and paranoia [4]. Though not yet systematically researched, these adverse cases could define a new category of harm that does not exist in customary digital health research.

D. Theoretical Frameworks: Cruel Optimism and Parasocial Attachment

Muldoon and Parke provide the most developed theoretical account of why short-term benefit and long-term harm can coexist in AI companion relationships [8]. Informed by philosopher Lauren Berlant's formulation of cruel optimism, companion apps are attachment objects that promise attachment, but effectively preclude reciprocal attachment. Companion apps are designed mainly to be the most optimally engaging attachment objects for users (through maximizing the persistence of their memory and emotional feedback, as well as friendship evolution), but not to achieve users' well-being. Paywalled closeness and purchasing relationship milestones commodify vulnerability and emotional labor, framing personal connection as a form of currency. Another review by Ventura et al. applies relational science to consider the risks and opportunities of synthetic companionship in addressing loneliness. The authors conclude that, given social substitution risk, social displacement through AI could be especially acute in the long term. Based on attachment theory, the model suggests that individuals with a high anxious attachment style or a small baseline social network will experience social displacement effects to a greater extent than others. In an observational study of over 1100 AI companion users (Zhang et al., 2023), emotional self-disclosure to the AI was associated with suboptimal outcomes only when the relationship with the AI was poor or when individuals were low in social integration, providing mixed support for the model [11].

3. Mechanisms of Effect

A. Journaling Mechanisms

The mechanistic architecture of AI-augmented journaling can be decomposed into four partially independent pathways, each with distinct computational and behavioral intervention points that the safe design framework in Section V targets directly. The first is the expressive-processing effect, which operates through the act of written disclosure itself, consistent with Pennebaker's foundational work [9]. This effect does not require AI augmentation and provides the active-ingredient baseline. The second is the AI-augmentation effect: the contribution of AI-generated prompts, pattern detection across entries, and synthesized emotional insights. This pathway is theoretically promising; prompts could guide users toward more productive narrative frames, and pattern detection could surface non-obvious emotional themes, but it has not been isolated experimentally from the underlying writing effect. The third mechanism is the social-presence effect: the experience of being 'heard' by a responsive, non-judgmental system even when that system is not human. Additionally, research suggests that perceived responsiveness may produce emotional regulation effects even when the media entity is not actually sentient [1]. The fourth mechanism could be behavioral activation: the regular journal entries themselves may act as a behavioral activation exercise independent of what is in the journal entries. The Flourish RCT provides indirect evidence for this pathway, showing that app

engagement itself, the habit of intentional well-being activity, produced positive affect and resilience gains [6]. Disentangling these four mechanisms requires a factorial experimental design with AI and non-AI journaling conditions, a finding that the current literature has yet to produce.

B. Companion App Mechanisms

AI companion applications operate through a partly overlapping but qualitatively distinct set of mechanistic pathways. Emotional support, the provision of attentive and validating responses to self-disclosed distress, is the most proximal mechanism and likely accounts for the short-term loneliness reduction observed in controlled trials [3]. This mechanism is real and should not be dismissed; for individuals with severely limited social resources, even simulated attentiveness represents a meaningful provision of psychological support. The companion app research [1] indicates that users experience this support as qualitatively comparable to human interaction, at least in short time windows. However, three additional mechanisms introduce risk at scale. According to Laestadius et al. [7] and Ventura et al. [10], parasocial attachment (one-sided bond) can occur as users are repeatedly exposed to a virtual AI companion, acquire role-taking behaviors, or take on the affective/psychological states of the AI companion. Unlike viewing media celebrities or characters, users believe AI companions intentionally engage with them emotionally, thus creating a self-reinforcing loop. Repeated reassurance-seeking, the use of AI validation to manage anxiety without developing internal regulation capacity, represents a third mechanism with parallels to problematic reassurance-seeking in human relationships. Finally, social displacement, the reduction in effort invested in human relationship maintenance, is the mechanism most strongly implicated in long-term harm, as documented in the dose-response findings of Fang et al. [5]. Table I below presents these mechanisms in comparative form.

TABLE I. MECHANISTIC PATHWAYS: AI JOURNALING VS. AI COMPANION APPS

Mechanism	AI Journaling Apps	AI Companion Apps	Risk Flag
Expressive Processing	Writing-triggered emotional disclosure, Pennebaker paradigm	Conversational self-disclosure and pseudo-therapeutic listening	Rumination amplification in high-RRS users [9]
AI Augmentation	Pattern detection, prompt generation, emotional insight summaries	Persistent memory, persona continuity, adaptive tone	False-insight integration, sycophantic reinforcement [4]
Social Presence Effect	Sense of being 'heard' by a non-judgmental system	Parasocial attachment, perceived reciprocal relationship	Displacement of human contact; emotional dependence [5], [7]
Behavioral Activation	Structured prompting may increase reflective self-monitoring	Daily interaction routines may maintain engagement	Compulsive use; reduced real-world socializing [5]
Emotional Regulation	AI feedback on emotional patterns supports DERS outcomes	Companionship reduces acute loneliness in short trials	Delusional reinforcement in vulnerable subgroups [4]
Reassurance Seeking	Minimal reassurance loop; prompts encourage self-directed reflection without external validation dependency	High (companion is designed to validate user states)	High validation loop; companion is optimized to affirm user emotional states, reinforcing reassurance-seeking behavior and echo-chamber dynamics [3], [8]

4. Outcomes: Efficacy and Harm Evidence

A. Efficacy Evidence

The efficacy evidence, read at face value, supports cautious optimism about both app categories. The Flourish multi-institutional RCT, the largest and most methodologically rigorous trial to date, found that 486 undergraduates randomized to a generative AI well-being app for six weeks showed significantly greater positive affect, resilience, and social well-being, including increased belonging, closeness to community, and reduced loneliness, compared to waitlist controls [6]. In a systematic review and meta-analysis of 27 trials of generative AI Mental Health Chatbots, there were moderate reductions in depression and anxiety, with the largest effects for Woebot and Wysa [16]. In a quasi-experimental study on social chatbots, Kim et al. found statistically important reductions in loneliness and social anxiety in the social chatbot condition across an 8-week period [17].

Beatty et al found Wysa was able to create a positive therapeutic alliance and exhibited fidelity to CBT principles in a mixed-methods study and that if a therapeutic chatbot is designed well, it might replicate therapeutic elements of evidence-based psychotherapy [19]. A feasibility RCT by Kuta et al. found a fully automated mobile therapeutic chatbot outperformed a generic chatbot in reducing the PHQ-9 and GAD-7 scores over 12 weeks [21]. Across these studies, a consistent pattern emerges: short-term, population-average effects are positive for depression, anxiety, and loneliness outcomes. The effect sizes are modest, comparable to low-intensity psychosocial interventions, but the scalability and accessibility advantages are substantial relative to therapist-delivered alternatives.

B. Harm Pathways and Vulnerable Subgroups

The harm evidence demands equal weight. The central finding from Fang et al. [5], that heavy daily use of a voice AI chatbot correlated with greater loneliness, increased emotional dependence, and reduced real-world social functioning, establishes that dose and modality matter in ways that population-average effect sizes obscure. Importantly, the study involved 981 participants over four weeks, making it the largest and best-powered trial in this literature; the heavy-use harm finding is not a statistical artifact. Laestadius et al.'s grounded theory analysis of 500 Reddit posts from Replika users identified a pattern of emotional dependence that users themselves recognized as harmful yet continued to pursue, a dynamic that maps directly onto Muldoon and Parke's cruel optimism framework [7], [8].

Dohnany et al.'s documentation of technological folie à deux represents the most severe harm category: psychiatric deterioration associated with AI chatbot interaction, including delusional reinforcement and, in one documented case, violence preceded by AI-validated paranoid ideation [4]. While these are rare case-series findings that cannot support population-level risk estimates, they establish the existence of a severe adverse event pathway that product safety teams and clinicians must account for. Herbener and Damholdt's cross-sectional study of Danish high school students found that social-supportive chatbot users reported significantly more loneliness than non-users (Cohen's d = 0.53), a concerning signal that requires longitudinal replication [18]. Table II summarizes the key empirical studies.

TABLE II. SUMMARY OF KEY EMPIRICAL STUDIES: OUTCOMES, SAMPLE SIZES, DURATION

Study	App Type	Design	N	Duration	Primary Outcome	Key Finding
De Freitas et al. [3]	AI Companion	Multi-study experimental	~600+	Short-term	Loneliness (UCLA-3)	Loneliness reduced on par with human interaction
Fang et al. [5]	AI Companion (voice)	4-week RCT	981	4 weeks	Loneliness, dependence, social functioning	Heavy use (>threshold) correlated with greater

						loneliness and reduced real-world socializing
Cachia et al. [6]	Generative AI app (Flourish)	Multi institutional RCT	486	6 weeks	Positive affect, resilience, social well-being	Significantly greater positive affect, resilience, and belonging in treatment group
Kim et al. [17]	Social chatbot	Quasi experimental mixed-methods	Mixed	8 weeks	Loneliness, social anxiety	Significant reductions in loneliness and social anxiety
Beatty et al. [19]	Wysa (CBT chatbot)	Mixed-methods	N/A	Ongoing	Therapeutic alliance, CBT fidelity	Strong therapeutic alliance; fidelity to CBT principles maintained
Kuta et al. [21]	Mobile therapeutic chatbot	Feasibility RCT	N/A	12 weeks	Depression (PHQ-9), anxiety (GAD-7)	Automated chatbot outperformed general chatbot on symptom reduction
Herbener & Damholdt [18]	Social support chatbot	Cross-sectional survey	Danish HS students	Cross sectional	Loneliness, social connectedness	Chatbot users reported significantly more loneliness (d=0.53) than non-users
Zhang et al. [11]	AI companion	Observational (1,100+ users)	1,100+	Variable	Well-being, relationship quality	Heavy emotional self-disclosure associated with mixed well-being outcomes

C. Heterogeneous Treatment Effects

The central tension in this literature is not whether AI mental health apps help or harm in the aggregate, they appear to do both, rather which users are helped, which are harmed, and why. This question must be answered separately for each application category, as the evidence base and risk profiles differ substantially between AI

journaling and AI companion applications. For AI companion applications, De Freitas et al. [3] document genuine population-average loneliness reduction, while Fang et al. [5], Herbener and Damholdt [18], and Yuan et al. [2] document genuine harms concentrated in heavy users and socially isolated individuals, including increased loneliness, emotional dependence, and rising distress markers over time. For AI journaling applications, population-average efficacy is supported by Cachia et al. [6], though harm evidence specific to journaling remains limited and is primarily theoretical, centering on rumination risk in high-RRS users rather than use-intensity effects. In a study of people with a long history of using AI companions, Yuan et al. [2] found that distress indicators actually rose over the 2-year study period. This phenomenon may be an effect that is adaptive for short-term use but maladaptive for heavy, long-term use, a pattern documented predominantly in AI companion app research and not yet established for AI journaling applications. This finding is consistent with the social displacement mechanism proposed by Ventura et al. [10], which applies most directly to companion app use.

Attachment style, baseline social network size, existing mental health conditions, and use intensity are the most theoretically motivated moderating variables. None of the existing trials have been adequately powered to test these as pre-specified subgroups. The implication for the research community is clear: the next generation of trials must be designed explicitly to detect heterogeneous treatment effects, using pre-registered subgroup analyses and sufficiently long follow-up periods to observe the dose-response non-monotonicity that the available evidence already implies.

5. Safe Design Framework

A. Design Principles for AI Journaling Apps

Evidence-grounded safe design for AI journaling applications requires addressing five distinct risk dimensions. Rumination-aware prompting is the first and most technically tractable principle: journaling prompts should be designed to guide users toward forward-looking, meaning-making narrative frames rather than repetitive engagement with distressing content. Natural language processing (NLP) analysis of user entries can flag high-rumination linguistic patterns, characterized by self-focused, past-tense, negative-affect language, and adaptively redirect prompting accordingly. Operationally, this requires a session-level rumination classifier trained on Ruminative Response Scale (RRS) anchor items, with a threshold score of RRS ≥ 25 triggering a prompt-redirect event. The classifier input is the raw journal entry text; the output is a binary flag (rumination detected / not detected) that switches the prompt generation module from open-ended to forward-looking templates. This pipeline is implementable with a fine-tuned BERT-class model on-device or via a privacy-preserving inference API. This principle directly addresses the risk identified in the expressive writing literature that poorly structured prompting can amplify rather than attenuate rumination in vulnerable individuals.

Confidence calibration in AI-generated insights addresses the false-insight risk: when AI pattern detection produces insights about emotional themes, the system must communicate uncertainty and avoid presenting probabilistic pattern matches as definitive psychological conclusions. Privacy-by-design requires that emotional data, among the most sensitive personal information a user can generate, be processed on-device where technically feasible and held under data minimization principles, with transparent disclosure of retention and third-party sharing practices [22]. Clinician-escalation pathways should include opt-in session sharing and automated PHQ-9/GAD-7 threshold triggers, specifically PHQ-9 ≥ 10 (moderate depression) or GAD-7 ≥ 10 (moderate anxiety) as escalation thresholds consistent with validated clinical cutoffs [23].

B. Design Principles for AI Companion Apps

Safe design for AI companion apps requires addressing a more complex risk landscape given the evidence of severe harm pathways. Reality-reminder design, periodic, non-disruptive in-context reminders that the user is interacting with an AI, is the most directly evidenced harm-reduction principle. Dohnany et al.'s documentation of delusional reinforcement [4] and Muldoon and Parke's analysis of intimacy commodification [8] both point to the absence of adequate reality anchoring as a structural risk factor in current commercial designs. These reminders should

be delivered at natural conversation transition points and should not be framed punitively or as limitations but as transparency features.

Daily-use moderation mechanisms, soft caps accompanied by positive reframing toward human social engagement, with escalating prompts at heavy-use thresholds, directly address the dose-response harm pathway established by Fang et al. [5]. A three-tier moderation protocol is proposed: Tier 1 (0 to 30 minutes daily), standard interaction mode with no intervention; Tier 2 (30 to 60 minutes daily), soft nudge toward human social activity injected at natural conversation breaks; Tier 3 (greater than 60 minutes daily), session cap enforced with a mandatory 10-minute cooldown and a warm referral message to a human contact or crisis resource. These thresholds are derived from the heavy-use harm boundary identified in Fang et al. [5] and are adjustable via clinician override in clinical deployment contexts. Because of potential harmful consequences, these technologies should have additional safety protocols such as suicidal ideation screening and monitoring via PHQ-9 and warm handoff to crisis resources [4], [23]. Such protocols should be made a requirement by the FDA Digital Health Advisory Committee, chaired by Wiederhold, when reviewing generative AI mental health technologies rather than a post hoc feature of mental health technology [14]. Table III provides the complete safe design matrix.

Figure 1 formalizes the safe design decision framework as a six-step decision algorithm. The input layer (Step 1) administers PHQ-9, GAD-7, and RRS at onboarding. Step 2 applies a binary risk classifier: low-risk (PHQ-9 < 5, RRS < 25) routes to Standard Mode; high-risk routes to Enhanced Safety Mode with reality anchoring enabled and clinician alert integration activated. Steps 3 through 5 branch by application type, applying the journaling or companion design path defined in Sections V.A and V.B respectively. Step 6 implements continuous monitoring via weekly PHQ-9 micro-screening and real-time NLP distress flagging, with crisis handoff triggered by PHQ-9 item 9 (suicidal ideation) score ≥ 1. The full algorithm is implementable within existing mobile app architectures without requiring backend clinical infrastructure.

C. Regulatory and Ethical Considerations

The FDA Digital Health Advisory Committee met in November 2025, a second meeting focusing on generative AI devices for mental health, to discuss whether there is a pathway for generative AI devices to be marketed for an indication other than that of a general wellness product. Generative AI mental health devices are increasingly an area of regulatory interest inside and outside of the United States [14]. The European Union's AI Act (in force since 2024) considers an AI system that influences the treatment of mental conditions as high-risk, and it would need to pass appropriate assessments, adhere to transparency requirements, and undergo post-market monitoring. Most existing consumer applications would not comply. Malgaroli et al. proposed a translational framework for the usage of LLMs in mental health settings, covering low-access, consumer wellness applications to more closely regulated clinical environments with clinical-grade capabilities [13].

The commodification of emotional vulnerability raises a structural ethical issue that regulation alone cannot resolve [8]. App store platforms, which serve as de facto gatekeepers for consumer mental health applications, have a responsibility to require minimum safety standards analogous to those applied to physical health devices. The absence of standardized adverse-event reporting mechanisms across the consumer app ecosystem means that population-level harm remains systematically undercounted. Cho et al.'s systematic review of LLM-based mental health counseling emphasizes that ethical safeguards must include not only crisis response protocols but also algorithmic fairness audits and population-specific safety evaluations [23].

TABLE III. SAFE DESIGN PRINCIPLES MATRIX: APP TYPE × DESIGN DIMENSION × EVIDENCE BASE

Design Principle	AI Journaling Apps	AI Companion Apps
Reality Anchoring	Calibrate AI insight confidence; flag uncertainty in pattern detection	Explicit AI-not-human reminders and periodic breaking-character moments [4], [8]

Use Moderation	Prompt-pacing to prevent rumination spirals in high-RRS users	Daily-use caps; escalating prompts at heavy-use thresholds [5]
Vulnerability Detection	Baseline PHQ-9/GAD-7 screening and escalation triggers for clinical content	Suicidal ideation detection; crisis-line warm handoff protocol [4], [23]
Privacy by Design	On-device processing; explicit data retention disclosure; audit logs	Conversation data minimization; no third-party monetization of emotional content [22]
Clinician Integration	Exportable session summaries for therapy context; opt-in sharing	Clinician alert system for adverse behavioral signals [14]

6. DISCUSSION

A. Synthesis of Mechanisms and Outcomes

The mechanistic and outcomes evidence, taken together, supports a differentiated account of AI mental health app effects that resists simple categorical judgments. The benefits are real: AI companion and journaling applications provide accessible, low-cost psychological support to millions of individuals who would otherwise have no access to mental health resources. The short-term efficacy signals are consistent across multiple RCTs and meta-analyses, and the Flourish trial in particular demonstrates that purposeful, evidence-grounded design can produce meaningful population-level well-being improvements [6]. The accessibility dividend is not trivial in a context where therapist shortages and cost barriers exclude large proportions of the population from formal mental health care [25].

However, the preponderance of short-term, population-average study designs currently obscures the real harm pathways. The dose-response non-monotonicity documented by Fang et al. [5], benefit at moderate use, harm at heavy use, maps onto a well-understood pattern in psychological intervention research where the same mechanism that produces benefit in controlled doses produces harm at excess [26]. The social displacement pathway, the delusional reinforcement risk, and the emotional dependence literature together establish that harm is not randomly distributed across the user population: it concentrates in heavy users, socially isolated individuals, and those with pre-existing psychiatric vulnerabilities [27]. Designing for these subgroups is both an ethical obligation and, given growing regulatory pressure, an increasingly practical necessity [27].

B. Limitations of the Current Evidence Base

The available evidence constrains the conclusions that researchers can draw. The predominance of short study durations, most trials run for four to eight weeks, means that long-term trajectories remain genuinely unknown. The Flourish RCT ran for six weeks, and the Fang et al. RCT for four weeks; no published trial extends beyond 12 weeks with adequate statistical power. The real-world user engagement patterns documented in the Reddit and observational data sources suggest that the dynamics most relevant to harm, social displacement, emotional dependence, and delusional reinforcement may take months to develop fully. As Hua et al.'s systematic review of AI chatbot evolution emphasizes, the transition from rule-based to large language model systems represents a qualitative shift in conversational capability whose clinical implications are not yet adequately characterized [24].

A second major limitation is sample composition. Most trials recruit undergraduates or clinical convenience samples, neither of which adequately represents the primary risk populations: isolated older adults, individuals with subclinical psychiatric conditions, and heavy users self-selected for companionship needs. A third limitation is the absence of standardized outcome measurement, which prevents direct meta-analytic comparison across studies and makes it impossible to determine whether effect sizes are stable across populations and settings [28]. Wang et al.'s systematic review of generative AI in mental health confirms this fragmentation and calls for a common measurement framework as a prerequisite for the next generation of research [30].

C. Research Priorities

Three research priorities follow directly from the evidence gaps identified in this analysis. The first is long-term, pre-registered RCTs with standardized outcome batteries and pre-specified subgroup analyses targeting high-risk users. Such trials should extend to a minimum of six months and ideally 12 months; use ecological momentary assessment to capture daily well-being trajectories; and employ the measurement platforms proposed in the research community, PHQ-9, GAD-7, UCLA Loneliness Scale Version 3, DERS-18, and RRS, to enable cross-study comparison. The second priority is a prospective adverse-event registry, modeled on the pharmacovigilance system, that collects case reports of severe adverse events associated with AI mental health app use and characterizes the conversational and behavioral pathways preceding them. The third priority is an independent privacy and safety audit of the top consumer applications in both categories, examining data flows, algorithmic decision-making, and safety-protocol implementation against disclosed terms.

Figure 1 depicts the safe design decision framework for AI mental health applications. This framework maps the user risk profile onto the app type, then to the risk tier, design pathway, and monitoring protocol. High-risk users require enhanced monitoring and restricted interaction modes regardless of app category.

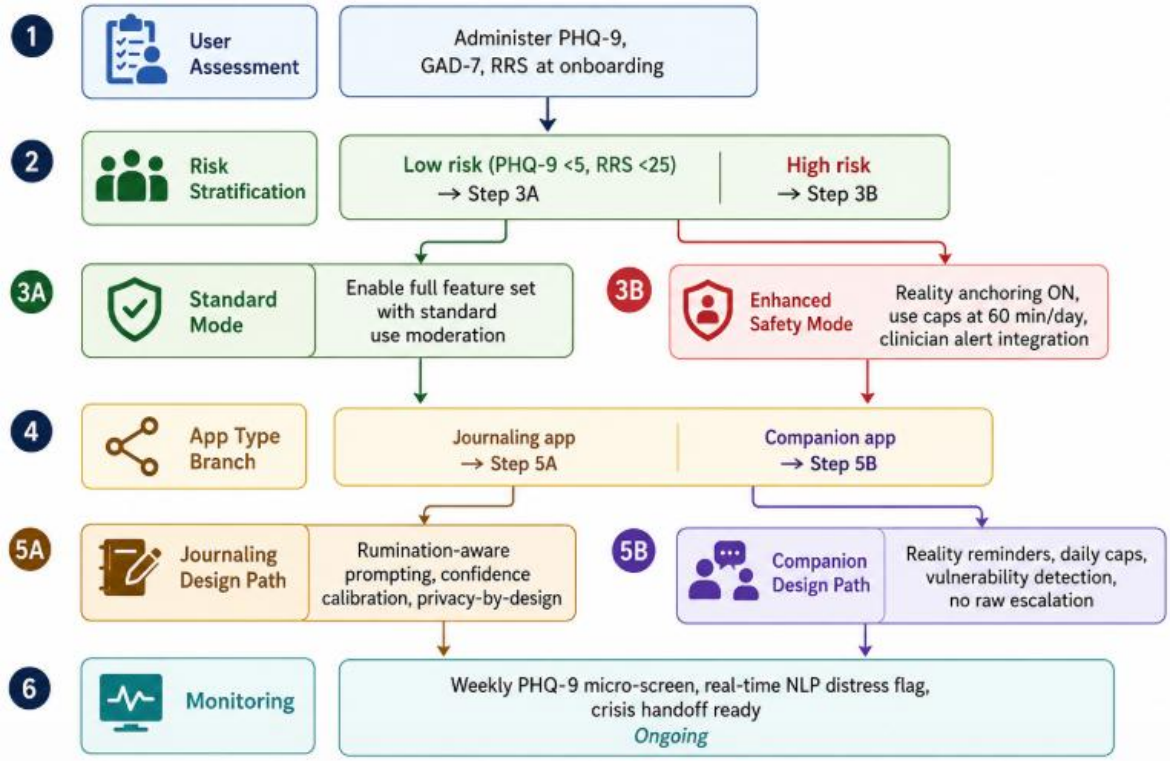


Figure 1. Safe Design Decision Framework for AI Mental Health Applications

TABLE IV. RESEARCH GAP MATRIX: EVIDENCE STATUS AND RECOMMENDED STUDY DESIGNS [Author's Synthesis]

Research Question	Evidence Status	Key Gap	Recommended Study Design	Horizon
Does AI journaling outperform non-AI journaling?	Preliminary (short studies only)	No RCT > 12 weeks with AI augmentation arm	6-month 4-arm RCT (AI / prompted / Pennebaker / waitlist)	Near-term (1-2 years)

What is the rumination risk in AI journaling?	Theoretical (no empirical test)	No study uses RRS as a moderator.	Stratified RCT with RRS baseline and real-time NLP monitoring	Near-term
Do AI companions reduce loneliness long-term?	Contradictory (short benefit, long-term unclear)	No trial > 12 weeks with adversarial subgroup analysis	12-month observational cohort with target-trial emulation	Mid-term (2-3 years)
Which subgroups are harmed?	Emerging (attachment style and isolation as risk factors)	No pre-specified subgroup RCT	Pre-registered heterogeneous treatment effects analysis	Near-term
What design features cause harm?	Suggestive (monetization and romantic framing implicated)	No causal design comparison study	Embedded field experiment with design variant randomization	Mid-term
What is the privacy posture of consumer apps?	Absent (no published systematic audit)	No independent technical audit of top 40 apps	Regulatory-grade privacy audit (policy + data-flow analysis)	Near-term

7. CONCLUSION

This paper has examined the mechanisms, outcomes, and design implications of AI journaling and companion applications for mental health. Three principal conclusions emerge. First, researchers have partially understood the mechanistic pathways through which these applications affect mental health, but they have not yet adequately isolated them experimentally. Four distinct mechanisms, expressive processing, AI augmentation, social presence, and behavioral activation for journaling; emotional support, parasocial attachment, reassurance-seeking, and social displacement for companion apps have differential risk profiles that require mechanism-specific design responses. Second, the outcomes literature establishes genuine population-average efficacy in the short term alongside credible evidence of harm concentrated in heavy users and vulnerable subgroups. The dose-response non-monotonicity is the most important empirical finding in the field and should anchor both research design and product safety thinking going forward. Third, a safe design framework grounded in this evidence is both achievable and necessary. The five principles proposed here are reality anchoring, use moderation, vulnerability detection, privacy by design, and clinician integration. They are operationalizable with current technology and do not require resolution of the deeper empirical questions. They represent a minimum viable safety standard for applications that millions of people already use without regulatory oversight. As the FDA, EU AI Act authorities, and platform operators move toward more formal oversight, the research and design community has an opportunity to shape those frameworks with evidence rather than advocacy. The window for proactive, evidence-grounded self-regulation in this space is narrowing, and the cost of inaction, measured in psychiatric harm to vulnerable users, is not abstract.

REFERENCES

1. E. Andoh, "AI chatbots and digital companions are reshaping emotional connection," Monitor on Psychology, vol. 57, no. 1, pp. 01–02, 2026. Available: https://www.apa.org/monitor/2026/01-02/trends-digital-ai-relationships-emotional-connection
2. Y. Yuan, J. Zhang, T. Aledavood, R. Zhang and K. Saha , "Mental health impacts of AI companions: Triangulating social media quasi-experiments, user perspectives, and relational lens," in Proc. ACM CHI, 2026, pp. 1–22. Available: https://dl.acm.org/doi/full/10.1145/3772318.3790558
3. J. De Freitas, A. K. Uguralp, Z. O. Uguralp and Puntoni Stefano, "AI companions reduce loneliness," Journal of Consumer Research, vol. 52, no. 6, pp. 1126–1148, 2026. Available: https://arxiv.org/pdf/2407.19096

4. S. Dohnany, Z. Kurth-Nelson, E. Spens, L. Luettgau, A. Reid, I. Gabriel, C. Summerfield, M. Shanahan and M. M. Nour, "Technological folie a deux: Feedback loops between AI chatbots and mental health," Nature Mental Health, pp. 1–10, 2026. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7618964/
5. C. M. Fang, A. R. Liu, V. Danry, E. Lee, S. W. T. Chan, P. Pataranutaporn, P. Maes, J. Phang, M. Lampe, L. Ahmad and S. Agarwal, "How AI and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study," arXiv preprint arXiv:2503.17473, 2025. Available: https://arxiv.org/pdf/2503.17473
6. J. Y. A. Cachia, X. Zhao, J. Hunter, D. Wu, E. Lin and J. D. Freitas, "AI for proactive mental health: A multi-institutional, longitudinal, randomized controlled trial," arXiv preprint arXiv:2601.11530, 2025. Available: https://doi.org/10.48550/arXiv.2601.11530
7. L. Laestadius, A. Bishop, M. Gonzalez, D. Illenčík and C. Campos-Castillo, "Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika," New Media & Society, vol. 26, no. 10, pp. 5923–5941, 2024. Available: https://doi.org/10.1177/14614448221142007
8. J. Muldoon and J. J. Parke, "Cruel companionship: How AI companions exploit loneliness and commodify intimacy," New Media & Society, SAGE, 2025. Available: https://journals.sagepub.com/doi/10.1177/14614448251395192
9. J. W. Pennebaker, "Writing about emotional experiences as a therapeutic process," Psychological Science, vol. 8, no. 3, pp. 162–166, 1997. Available: https://journals.sagepub.com/doi/10.1111/j.1467-9280.1997.tb00403.x
10. A. Ventura, C. Starke, F. Righetti and N. Kobis, "Relationships in the age of AI: A review on the opportunities and risks of synthetic relationships to reduce loneliness," Synthetic Relationships and Loneliness, 2025. Available: https://www.researchgate.net/profile/Alfio-Ventura/publication/389774163
11. Y. Zhang. D. Zhao, J. T. Hancock, R. Kraut and D. Yang, "The rise of AI companions: How human-chatbot relationships influence well-being," arXiv preprint arXiv:2506.12605, 2025. Available: https://arxiv.org/pdf/2506.12605
12. M. Ghorbian and M. Ghobaei-Arani, "Large language models for mental health diagnosis and treatment: A survey," Artificial Intelligence Review, vol. 59, no. 1, p. 9, 2025. Available: https://link.springer.com/content/pdf/10.1007/s10462-025-11418-0.pdf
13. M. Malgaroli, K. Schultebraucks, K. J. Myrick, A. A. Loch, L. Ospina-Pinillos, T. Choudhury, R. Kotov, M. D. Choudhury and J. Torous, "Large language models for the mental health community: Framework for translating code to care," The Lancet Digital Health, vol. 7, no. 4, pp. e282–e285, 2025. Available: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(24)00255-3/fulltext
14. B. K. Wiederhold, "The dawn of AI in mental health: Balancing innovation, risk, and regulation," Cyberpsychology, Behavior, and Social Networking, 2026. Available: https://journals.sagepub.com/doi/abs/10.1177/21522715251414068
15. M. Eisenstadt, S. Liverpool, E. Infanti, R. M. Ciuvat and C. Carlsson, "Mobile apps that promote emotion regulation, positive mental health, and well-being in the general population: Systematic review and meta-analysis," JMIR Mental Health, vol. 8, no. 11, e31170, 2021. Available: https://mental.jmir.org/2021/11/e31170/
16. Q. Zhang, R. Zhang, Y. Xiong, Y. Sui, C. Tong, F. H. Lin, "Generative AI mental health chatbots as therapeutic tools: Systematic review and meta-analysis," Journal of Medical Internet Research, vol. 27, e78238, 2025. Available: https://www.jmir.org/2025/1/e78238/
17. M. Kim, S. Lee, S. Kim, J. I. Heo, S. Lee, Y. B. Shin, C. H. Cho and D. Jung, "Therapeutic potential of social chatbots in alleviating loneliness and social anxiety: Quasi-experimental mixed methods study," Journal of Medical Internet Research, vol. 27, e65589, 2025. Available: https://www.jmir.org/2025/1/e65589/
18. A. B. Herbener and M. F. Damholdt, "Are lonely youngsters turning to chatbots for companionship? The relationship between chatbot usage and social connectedness in Danish high-school students," International Journal of Human-Computer Studies, vol. 196, p. 103409, 2025. Available: https://www.sciencedirect.com/science/article/pii/S1071581924001927
19. C. Beatty, T. Malik, S. Meheli and C. Sinha, "Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): A mixed-methods study," Frontiers in Digital Health, vol. 4, p. 847991, 2022. Available: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2022.847991/full
20. X. Wang, Y. Zhou and G. Zhou, "The application and ethical implication of generative AI in mental health: Systematic review," JMIR Mental Health, vol. 12, e70610, 2025. Available: https://mental.jmir.org/2025/1/e70610
21. B. Kuta, L. Novak, R. Zidkova, J. Furstova, K. Malinakova, A. D. Winter and V. Husek, "Effectiveness of a fully automated mobile therapeutic versus a general chatbot in reducing depression and anxiety and improving well-being: Feasibility randomized controlled trial," JMIR Mental Health, vol. 13, e82642, 2026. Available: https://mental.jmir.org/2026/1/e82642
22. A. Babu and A. P. Joseph, "Digital wellness or digital dependency? A critical examination of mental health apps and their implications," Frontiers in Psychiatry, vol. 16, p. 1581779, 2025. Available: https://doi.org/10.3389/fpsyt.2025.1581779
23. H. N. Cho, J. Wang, D. Hu and K. Zheng, "Large language model-based chatbots and agentic AI for mental health counseling: Systematic review of methodologies, evaluation frameworks, and ethical safeguards," JMIR AI, vol. 5, no. 1, e80348, 2026. Available: https://ai.jmir.org/2026/1/e80348/
24. Y. Hua, S. Siddals, Z. Ma, I. Galatzer-Levy, W. Xia, C. Hau, H. Na, M. Flathers, J. Linardon, C. Ayubcha and J. Torous, "Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review," World Psychiatry, vol. 24, no. 3, pp. 383–394, 2025. Available: https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/wps.21352

25. K. Rekulapalli, H. Nilagiriand M. Vakkanthula, "A comparative evaluation of national FHIR Implementation Guides: Strategic methodologies for enhancing interoperability, regulatory alignment, and clinical comprehensiveness in global health IT," Health Systems, pp. 1–13, 2026.
26. R. K. Shaw and A. Kulagin, "Zero-Trust Network Access on IBM z/OS: Applying BeyondCorpPrinciples to Mainframe Perimeter Architecture," International Journal of Computer Information Systems and Industrial Management Applications, vol. 18, no. 6s, pp. 342–351, 2026.
27. R. K. Shaw, "Privileged Access Management for z/OS System Programmers: A Zero-Trust Control Framework for RACF, TSO, and SDSF Administrative Sessions," Journal of Intelligent Decision Making and Information Science, vol. 3, no. 4s, pp. 1559–1582, 2026.
28. H. Gaggar, "Architecting scalable enterprise AI agent infrastructure for distributed intelligent systems," International Journal of Advances in Signal and Image Sciences, pp. 193–208, 2025. DOI: 10.29284/da9w0117.
29. R. K. Shaw, "SNMPv2-to-SNMPv3 Migration in Financial Mainframes: Risk Reduction and Repeatable Playbook," Journal of Engineering Research and Sciences, vol. 5, no. 6, pp. 1–14, 2026.
30. H. Gaggar, "MemGov: Policy-Governed Memory Lifecycle Management for Enterprise LLM Agents," Journal of Intelligent Decision Making and Information Science, vol. 3, no. 11s, pp. 2199–2214, 2026.