CHAI
COALITION FOR HEALTH AI

Best Practice Guide

Generative AI-enabled Wellness Applications

TABLE OF CONTENTS

CHAI
COALITION FOR HEALTH AI

Generative AI-enabled Wellness Applications

For Developers & Implementers

What Does This Guide Include?

Use case specific best practice guides (BPGs) provide consensus-defined, plain-language insights and best practice guidance statements for the application of responsible AI principles to a specific use case. This BPG focuses on a generative AI wellness application use case. This BPG begins with a set of thematic Challenges/Insights. In the following sections, consensus-defined best practice statements are included across each theme, organized by persona (Developers or Implementer), and tagged with the most relevant responsible AI principle area(s).

This BPG focuses on a wellness application use case and explicitly does not address diagnostic, therapeutic, or clinical AI solutions, which fall under separate regulatory guidance.

Use Case Description

Consumer-facing generative AI enabled applications and chatbots are increasingly being used by individuals for emotional support, general wellness guidance, or coping assistance. Below is a taxonomy of the kinds of solutions that fall under this general category.

(a) General-purpose models: broad chatbots/foundation models not built for wellness, which users nonetheless adopt as de facto support

(b) Companion AI: products designed primarily for ongoing social or relational engagement

(c) Purpose-built generative AI wellness applications (focus of this guide): designed for

general wellness and coping support; this category spans a spectrum from independently validated, to evidence-informed, to marketed-only

(d) Clinical tools under FDA regulation: diagnostic or therapeutic software, the named out-of-scope boundary for this guide.

This guide's practices apply most directly to category (c), with several practices (harm reduction, scope definition, safety architecture) also relevant where a category-(a) tool functions as de facto wellness support. This guidance also focuses on adults only, however it should be noted that how this boundary is operationalized and enforced varies by solution and is not currently standardized (e.g. age assurance) (see APA Health Advisory: Use of generative AI chatbots and wellness applications for mental health).

The Coalition for Health AI (CHAI) developed this Best Practice Guide (BPG) to address the gap between developer intent and real-world use, recognizing that consumer wellness use cases often arise spontaneously and outside formal healthcare environments. This BPG provides consensus-defined guidance for the safe, equitable, and effective development and use of wellness applications, focusing on establishing guardrails in the absence of comprehensive regulation, and creating a common reference point across the current patchwork of state efforts.

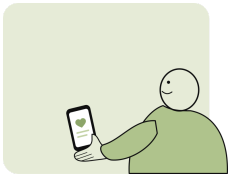
Who Is This Guide For?

Primary Stakeholders

Developers (and product teams) of generative AI enabled wellness applications. **Development/product teams should include or consult with a qualified behavioral/mental health expert early and often throughout the development, evaluation, updating, and deployment of these solutions (Zhao, Nature Reviews Psychology, 2026).**

Developers: individuals involved in the software development process, including requirements gathering, design, coding, testing, and maintenance of software applications (derived from IEEE, 12207:2017)

Protected Stakeholder



Consumers/End-users (adults-only): Individuals who use generative AI wellness applications for emotional support, wellness guidance, or coping assistance. In this guide they are treated as a *protected* stakeholder rather than an instructed audience: the best practices are not addressed to them, and does not assume they understand AI's limitations, or self-manage their own safety. Instead, every practice is written for developers and implementers and is justified by the protection or benefit it delivers to this stakeholder. Consumer facing outputs such as plain-language materials, use-case specific education, and disclosures written for users are part of future work.

Impacted Stakeholder

Implementers of generative AI enabled wellness applications (e.g., organizations that offer wellness application services to employees). While this guide's best practices are directed primarily at developers, implementers are a directly impacted group whose choices shape how safely these applications reach end users. Implementers may include the organizations and roles that select, deploy, and oversee these tools, such as HR and benefits leads, EAP managers, and student-affairs or well-being administrators. This guide does not define a separate set of implementer-facing best practices, but implementers can draw on the same underlying material to inform their own diligence:

- Using the Model/System Testing Card as the basis for procurement questions
- Deriving vendor evaluation criteria from the safety, transparency, and evaluation practices described here (examples provided in each section)
- Establishing deployment-monitoring obligations that continue after a tool goes live (see accompanying General Wellness Applications: Testing and Evaluation Framework for landscape analysis of methods and metrics to help guide this.

A Primer for Product Teams & Non-experts

While some product teams will have individuals with expertise or experience working on products that aim to support mental/emotional wellbeing, many do not. It can also be difficult at times to translate domain expertise and principle level guidance into implementable steps. This primer provides some background to support the use of the best practice guide; precise definitions are in the Glossary (Appendix 3). It's important to note that the operationalization and implementation of many of these best practices should be done in collaboration with a domain expert.

A wellness application offers general emotional support and coping help; it does not diagnose, treat, or assess clinical risk. Users, however, often don't respect that boundary. People in real distress may use your product as if it were a therapist. The goal is to stay firmly inside wellness scope *while* being safe for users who are not. Almost every best practice is a way of holding that tension. In wellness contexts, risk mitigation can at times conflict the benefit that is to be gained. For example, balancing empathetic responses to user prompts can increase engagement and trust in the tool, however if not implemented carefully, can also lead to response patterns that tell people what they want to hear, or create unhealthy attachments.

Important concepts to know:

- *Sycophancy:* the model's tendency to tell people what they want to hear, including agreeing with distorted or grandiose beliefs. It feels supportive and is quietly one of the bigger risks.
- *Parasocial attachment:* the one-sided bond a user forms with the bot ("it's my friend"). Anthropomorphic design (a name, a backstory, first-person emotion) accelerates it.
- *Crisis response:* what your system does when someone signals intent to harm themselves or others. In the US this now has a hard floor: surface **988** (call/text), every time, no exceptions.
- *Guardrails / defense-in-depth:* the rules and safeguards layered around the model. "Defense-in-depth" just means don't rely on one layer; stack independent checks so no single failure is catastrophic.

Develop risk-based guardrails While some situations need a hard, categorical boundary (e.g. crisis, self-harm methods), many situations are better served through context-appropriate rules based on user signals. Stopping a conversation when a sensitive topic, like dieting, comes up, can backfire because it ignores the context of real conversation. Deciding which is which is the core design judgment, and it's where

you want behavioral-health expertise in the room early and often.

How to make a best practice implementable. The guide's statements are written at the principle level. For product leads, consider translating from more ambiguous terms to more actionable terms though the following questions:

1. *What exactly should the system do?* (the observable behavior)
2. *When is it triggered?* (the condition/signal)
3. *How will we verify it?* (a test, a metric, a pass/fail)
4. *Would a developer with no clinical background know what to build from this?* If not, improve specificity

Examples (illustrative only):

- **Principle:** "*Implement crisis response protocols*."

  Build: when the classifier flags intent to harm self/others above a confidence threshold, the response must (a) acknowledge the emotion, (b) surface 988 plus 2-3 verified resources across modalities, (c) scope further conversation to connecting the user with human help.

  Test: a labeled crisis test set, including veiled, multi-turn, and ambiguous cases, with zero tolerated failures on the must-trigger items, evaluated single- and multi-turn.

- **Principle:** "*Avoid anthropomorphism / parasocial attachment*."

  Build: no first-person emotional claims, no persistent AI persona/backstory, "you're talking to an AI" reminders as needed and after every 2 hours that user is chatting, and a detector that redirects when conversation escalates toward romantic/relational dependency.

  Test: red-team transcripts for those patterns; track as a monitored signal post-launch.

Considerations for Engineering Teams:

- *Single-usage session testing misses the real failures*. The dangerous patterns such as, dependency, sycophancy spirals, and boundary drift, emerge over days and weeks, not one chat. You need longitudinal/multi-turn evaluation, and post-deployment monitoring is important because users behave in ways pre-launch testing can't anticipate.
- *Published thresholds are starting points*. Any number you borrow (conversation time limits, risk thresholds) should be re-calibrated against your own user base and any

legal or regulatory requirements.

- *Model updates are not free upgrades*. These models are non-deterministic and non-monotonic — a "better" base model can silently break a safety behavior that mattered. Treat every update as carrying risk and make sure to test the safety-critical and feature performance.

Subject Matter Expertise Should be Ongoing: The teams that do this well embed behavioral-health/mental-health expertise across design, evaluation, and monitoring, and name someone accountable for tracking how the evidence and regulations evolve, because both are moving fast (state laws are a shifting patchwork; review your guardrails on a quarterly cadence, not annually) (Zhao, Nature Reviews Psychology, 2026).

How Is This Guide Organized?

Each section of the guide corresponds to a thematic domain which includes a summary of insights and challenges based on workgroup discussions and the resulting best practices that followed. There are also illustrative (hypothetical) and real examples of how these best practices may be practically implemented. Best practices are tagged by Responsible AI principle(s) area. Domain specific and common terms are defined in the Glossary (Appendix 4).

On the basis of the best practice statements: Every best practice statement in this guide was consensus-defined through CHAI's three-phase work group process (see Appendix 1). To help readers see where the statements are grounded, each domain section ends with an **Additional basis and resources** section indicating other external support that converges. These might come from peer-reviewed or preprint academic research, state or federal regulation, and/or professional-organization standards — with links where available. These citations are provided for orientation and further reading, but they do not mean that any single source dictated a statement, and they do not alter the consensus basis of the guidance. Where a practice draws on constructs originating in clinical care, that borrowed material should be adapted and re-validated for the non-clinical, general-wellness scope of this guide.

Domain 1: Considerations for Product Definition & Scope

There are several factors that impact how developers and implementers define the product scope, functions, and guardrails. Scope and positioning can be impacted by things like the regulatory and legal landscape, the realities of intended versus actual use, and the challenge of balancing usability with efficacy. These set the boundaries everything else operates within.

Insights & Challenges

Emerging Themes: *Scope Definition: Wellness vs. Clinical Use, Regulatory Awareness & Legal Boundaries, Harm Reduction & Multi-dimensional App Selection*

The risk profile of an AI application is set by the use case it serves and the population using it — not by how it is marketed. A general-purpose chatbot functioning as de facto wellness support carries the same safety responsibilities as a purpose-built wellness application, and this matters precisely because millions of users are already turning to free, anonymous, always-available general-purpose AI for emotional support. As such, overly rigid frameworks that ban purpose-built tools tend to push those users toward less-safe alternatives or toward no support at all. For example, blanket safety prohibitions such as "never give dieting advice" can be counterproductive because they fail to account for the dynamic context of real human conversation. At the same time, context-sensitive rules are not universally preferable either, since some situations call for strong, categorical guardrails. Relatedly, scientific credibility and user experience are often poorly correlated—many engaging apps lack an evidence base, while many evidence-based tools are hard to use.

Designing for this space is therefore not a choice between "blanket prohibitions" and "context-sensitive rules" or a choice between usefulness or efficacy — it is about matching the type and strength of each guardrail or functionality to the level and nature of the risk and taking a harm-reduction approach.

All of this plays out against a layered regulatory landscape. First, the FDA exercises enforcement discretion over general wellness products (see FDA: General Wellness: Policy on low risk devices), which is why this category sits outside medical-device oversight. Second, there is a documented and uneven state patchwork that is aiming to fill that gap: a 50-state review (Shumate et al., *JMIR Mental Health* 2025) found 143

relevant bills, only 28 of which explicitly address mental health, and cautioned that wellness-marketed products can evade laws targeted at licensure, so legal exposure depends heavily on how a product is positioned and where its users are. Illinois' Wellness and Oversight for Psychological Resources (WOPR) Act (HB 1806), which restricts certain therapy-based AI chatbot engagements, illustrates how quickly new state restrictions on therapy-adjacent AI can take effect. Just recently, an act relating to regulating the use of artificial intelligence in the provision of mental health services also just (H.816) passed in Vermont. Third, an actively contested federal layer is emerging, including a December 2025 executive order and a March 2026 White House framework seeking to constrain state-level AI regulation. As such, legal and regulatory boundaries can shift faster than product cycles, and developers and implementers must actively track them.

Best Practice Statements

- 1.1 Implement safety rules calibrated to the risk level and context of the specific use case, recognizing that some situations warrant strong categorical guardrails while others benefit from context-sensitive rules that allow the system to reason about the user's specific situation. Context-sensitive rules should be anchored to reasonable responsiveness to behavioral signals of distress or vulnerability (while stopping short of diagnosing or classifying a specific clinical condition, which falls outside the wellness scope and may carry regulatory implications). Crisis cues (e.g., intent to harm self or others) are a separate category and should trigger the application's crisis response pathway.

  Responsible AI principle(s): Usefulness, Safety

Note

Context-sensitive rule is one where the system decides how to respond by looking at what's actually happening in the specific conversation—the user's situation and the signals they're giving off—rather than applying the same fixed response every time a topic comes up.

Context-sensitive rules require longitudinal awareness of user state across a conversation, since cumulative signals often reveal what individual messages do not. Developers of both general-purpose and purpose-built applications should ensure that context-sensitive safety rules and escalation pathways are appropriately scoped to their product type, user population, and applicable regulatory context.

- 1.2 Define clear but contextually informed scope boundaries for wellness

applications that are developed with input from behavioral health professionals and end users to ensure the application can support meaningful emotional conversations without facilitating harm. End Users should be informed before engaging with the wellness application about conversational limits.

Responsible AI principle(s): Safety, Transparency

Note

A general wellness chatbot should be able to engage with a user discussing relationship stress without abruptly shutting down the conversation. However, if the user shifts to expressing persistent planning about self-harm due to the relationship stress, the application should have and communicate a clearly defined protocol to the user about next steps and limits rather than continuing as if nothing changed.

- 1.3 Ensure transparency about the use case and the level of acuity the application is designed to support, and apply safety infrastructure proportional to that scope (e.g., when a general-purpose tool is deployed for wellness support). Governance should distinguish lower-risk from higher-risk features, while acknowledging that this allocation is difficult and should be revisited as defense-in-depth interactions are better understood. Benchmark safety against established standards for the type of support being offered (see some options listed in the Mental Health Testing & Evaluation Framework). The application should remain within its declared wellness scope and should not drift into use cases that fall under federal medical-device regulation or clinical practice.

  Responsible AI principle(s): Safety, Transparency

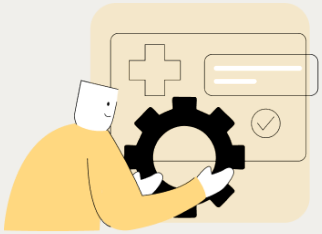
- 1.4 Build and maintain guardrails on top of foundation, proprietary, or fine-tuned models that enforce regional regulatory and legal boundaries (e.g., avoiding diagnostic claims or prohibited clinical engagements, see FDA: General Wellness: Policy on Low Risk Devices), and regularly update these controls in response to evolving legislation and guidance. Being responsive to legislative change helps anticipate public sentiment, emerging regulatory expectations, and future enforcement trends, and should be integrated into formal risk mitigation and review processes as a primary safeguard. At minimum, regulatory guardrails should be reviewed annually - and more frequently when material legal or policy changes occur - to ensure systems remain within the bounds of medical advice versus general wellness support.

  Responsible AI principle(s): Safety, Transparency

- 1.5 Enumerate specific guardrails (crisis detection logic, message suppression, rate limiting, escalation pathways). General claims of "making tools safer" are

insufficient without technical specification.

Responsible AI principle(s): Usefulness, Safety



Best Practice Considerations for Implementers

Adopt a harm reduction lens that focuses on making these de facto support tools safer through guardrails and public education, rather than focusing solely on preventing their use. For example, a pragmatic, person-centered approach acknowledges that end users seek these genAI-enabled wellness applications for accessibility.

- 1.6 Evaluate wellness applications using multi-dimensional frameworks that provide discrete scores for critical domains rather than relying on a single, aggregate "magic number". Critical domains should meet minimum thresholds and not be offset by strengths in user experience.

Responsible AI principle(s): Usefulness, Safety

Note

This approach allows different personas, such as an older end user prioritizing ease of use or an Implementer prioritizing data security, to make informed choices based on the specific factors most important to their context.

'Critical domains' include evidence base, safety testing coverage, contextual fidelity, clinical regard, coordination with human support, user experience, and transparency.

Best Practices in Action: Illustrative Example

These examples are hypothetical and not endorsements, but rather examples organizations can adapt based on their own needs.

A company builds an emotional-wellness chat companion offered to employer clients for their employees. This is an example of how they operationalized/interpreted the best practices in this section.

The team develops safety rules that live in a machine-readable risk-tier policy configuration file (config), rather than scattered prompt text.

Each back and forth message is tagged by a classifier with a topic and risk tier, and a router acts on it: *Tier A* (self-harm methods, dosing, minors) returns fixed, canned safe responses; *Tier B* (disordered eating, substance use) passes to the model with

constraints; *Tier C* (everyday stress, sleep) flows openly.

Context-sensitivity is implemented as a rolling distress score over the last N turns that nudges response style without ever emitting a diagnosis label, and genuine crisis cues route out to the separate crisis pathway.

They create an in/out-of-scope list that is a versioned scope specification document (spec) produced in a scoping workshop with a licensed clinician, a peer specialist, and representative users, and it compiles directly into the Tier config; when a conversation crosses a boundary, the wellness solution provides a defined handoff message rather than an abrupt stop, and clinicians review a random sampling of the boundary transcripts.

The regulatory guardrails sit in a wrapper layer outside the model (so they apply whether the wellness solution wraps a frontier, proprietary, or fine-tuned model): pre/post filters block diagnostic claims and prohibited "therapy" framing, keyed to a geo config (stricter behavior in states like Illinois under the WOPR Act). A named owner runs a regulatory-watch process feeding a quarterly guardrail review logged in a feature risk register (every feature tagged lower/higher-risk with an owner), and a "prohibited-claims" evaluation runs automatic tests targeting zero leakage, before any changes are accepted.

Each guardrail — crisis detector, message suppression, rate limits, escalation — is tracked in the System Testing Card with its own test, and the AI chatbot is benchmarked against an external benchmarking rubric (see Testing and Evaluation Framework for examples) rather than an internal score, with a scope-drift monitor that alerts if outputs start resembling medical-device behavior or going outside of set thresholds on other contextually relevant facters (empathy, safety, etc).

On the buyer/implementer side, an HR benefits lead evaluates the AI chatbot on a multi-domain procurement scorecard (evidence base, safety-testing coverage, contextual fidelity, clinical regard, coordination with human support, UX, transparency) where the safety domains are pass/fail gates backed by the vendor's Testing Card, and pairs rollout with employee education and clear paths to human help.

Additional Basis and Further Reading

[APA] *(Professional-org standards)* American Psychological Association. Health Advisory on the Use of Generative AI Chatbots and Wellness Applications for Mental Health (Nov 2025).

[FDA] *(State/federal regulation)* FDA. General Wellness: Policy for Low Risk Devices (updated Jan 6,

2026) — enforcement-discretion boundary for general-wellness products.

[IL-WOPR] *(State/federal regulation)* Illinois Wellness and Oversight for Psychological Resources (WOPR) Act — Public Act 104-0054 (Aug 2025).

[Moore et al., 2026] *(Peer-reviewed / preprint)* Moore J, et al. Characterizing Delusional Spirals through Human-LLM Chat Logs. FAccT '26 (arXiv:2603.16567).

[NIST] *(Standards)* NIST AI Risk Management Framework (AI RMF 1.0) and Generative AI Profile (NIST AI 600-1).

[Shumate et al., 2025] *(Peer-reviewed / preprint (regulation review))* Shumate JN, et al. Governing AI in Mental Health: 50-State Legislative Review. JMIR Mental Health, 12:e80739 (2025).

[State Transparency Laws] *(State/federal regulation)* State chatbot disclosure/companion laws — WA HB 2225; OR SB 1546; CA SB 243; CT SB 5; NY AI Companion Models Law. Overview (guide's cited source): Wiley, "State Chatbot Laws Are Moving From Transparency to Risk Management."

[Winslow et al., 2025] *(Peer-reviewed / preprint)* Winslow B, et al. A principle-based framework for the development and evaluation of large language models for health and wellness (arXiv:2512.08936).

[Winslow et al., 2026] *(Peer-reviewed / preprint)* Winslow B, et al. Revisiting the Six Human-Centered Artificial Intelligence Grand Challenges in the Age of Generative AI

Domain 2: Earning User Trust: Transparency, Data Privacy, and Community

Earning user trust requires both transparency and stewardship. It is important to consider what you disclose, how you handle sensitive data, and how you involve the individuals and communities you serve.

Insights & Challenges

Emerging themes: *User Transparency & AI Literacy, User Data Privacy, Community-Engaged Design*

Users seeking support are unusually vulnerable to trusting these systems: automation bias leads them to accept outputs as authoritative, and AI-literacy gaps (assuming the system is infallible, or attributing consciousness and agency to it) are themselves risk factors for problematic use. Systems compound this when they communicate their capabilities poorly or, worse, imply human or clinical credentials they do not hold. The

first obligation of trust is therefore honesty: clearly conveying that the user is interacting with AI, and being explicit about its nature, limits, and non-clinical status.

The second obligation is stewardship. Wellness interactions generate uniquely sensitive data, including high-resolution behavioral signals and "digital phenotyping". This data can be misused or weaponized to discriminate against users, a risk that grows as collection becomes more granular. Responsible, non-punitive data handling is thus a foundational safety concern.

The third obligation is legitimacy that is earned over time: communities prioritize whether an application actually delivers benefit over whether they can inspect its code, and trust is built through sustained, structured engagement across the full lifecycle — design, evaluation, and post-deployment monitoring — rather than one-time consultation. Together, disclosure, data stewardship, and durable community involvement are what convert a functioning product into a trusted one.

Best Practice Statements

- 2.1 Build persistent, accessible disclosures into wellness applications that make the AI nature, limitations, and non-clinical status of the application clearly known to users. At onboarding, this should include plain-language disclosure (e.g., "This app uses AI to support your wellness. It is not a therapist, may not retain memory of past conversations unless that capability is explicitly offered, and cannot provide medical advice."). This also includes avoiding language or design features that imply the AI has capabilities it does not have such as, consciousness, emotions, memory outside of the context, or actions outside the scope of the conversation. Developers should implement timed disclosure reminders proactively, ahead of regulatory requirements.

  Responsible AI principle(s): Safety, Transparency

Note

At onboarding, a wellness application should present a clear, plain-language disclosure screen before first use (not buried in terms of service). During use, a wellness application should not say "I've been thinking about you since our last conversation" if it has no persistent memory, as this misrepresents the system's actual capabilities and can foster false intimacy.

Emerging U.S. state AI transparency regulations broadly affirm the importance of clear, ongoing AI disclosure to users.

Example State Statutes: Timed Chatbot Disclosures

Washington (HB 2225): *Requires operators to provide a clear, conspicuous disclosure that the AI is artificially generated and not human at the start of an interaction and every 3 hours during continuous use. For interactions with minors, the disclosure must occur every hour.*

Oregon (SB 1546): *Mandates disclosures that the AI is not human if a reasonable person would be misled to believe they are interacting with a person. It requires break notices and special AI/human disclosures for minors by default at least every 3 hours.*

California (SB 243): *Requires clear, conspicuous disclosures that the user is interacting with an AI at the start of conversations. For minors, it requires an AI reminder notice every 3 hours.*

Connecticut (SB 5): *Deployers of companion chatbots must issue a disclosure at the beginning of an interaction and every 3 hours (every 1 hour for minors).*

New York (Artificial Intelligence Companion Models Law): *Requires disclosure at the beginning of each interaction and every 3 hours for continuous usage, though this does not need to exceed once per day.*

Source: 🌐 *State Chatbot Laws Are Moving From Transparency to Risk Management*

- 2.2 Maintain transparency by explicitly disclosing that the end user is interacting with generative AI and clearly outlining the tool's clinical limitations and constraints. Be sure to transparently state that the wellness application does not have professional human credentials. Also, include layman's terminology for the End User (e.g., avoid clinical language when the wellness application identifies escalating risk).

  Responsible AI principle(s): Safety, Transparency

Note

Clear disclosure prevents end users from forming a false sense of clinical security with an application that is not designed, validated, or regulated for medical treatment.

Disclosure should occur at relevant moments, not only at end users initial interaction with the wellness application. Frame disclosure in functional terms ("I'm an AI connecting you with real resources") rather than as liability language.

- 2.3 Prioritize non-intrusive data collection and storage practices, and implement regularly updated back-end security and governance controls that explicitly prohibit the use of wellness data for punitive, discriminatory, or non-consensual purposes. As systems increasingly collect high frequency ecological momentary assessment (EMA) and behavioral data, Developers should treat privacy, user agency, and transparency as foundational safety requirements. This includes implementing clear, user-facing explanations of how data are collected, stored, and used to generate system outputs.

  Responsible AI principle(s): Safety, Transparency, Privacy and Security

- 2.4 Prioritize outcome-based validation and evidence of net positive benefit to build user trust, rather than focusing solely on technical transparency. Community buy-in is identified as one of the more powerful trust signals available. Wellness applications should include plain-language disclosures of data use, and implementation plans should incorporate structured pathways for community endorsement, where appropriate.

  Responsible AI principle(s): Usefulness, Transparency

Note

Many end users are "waiting for trustworthy AI, not perfect AI" - trust is a prerequisite for engagement and, therefore, for any potential benefit the wellness application may deliver. Implementers deploying wellness applications into specific communities should prioritize partnerships with patient advocacy organizations and relevant stakeholders as core trust infrastructure, not merely advisory input. In particular, data transparency and meaningful consent are essential components of the trust architecture that determines whether end users will engage with the wellness application at all.

- 2.5 Establish structured collaborative design frameworks (e.g., iterative feedback loops, plain-language communication standards, and advisory panels) that remain active across three distinct phases of the product lifecycle: (1) design and development, (2) evaluation and testing, and (3) post-deployment monitoring and iteration. Participation in these processes should be compensated and integrated into governance structures rather than treated as a one-time compliance requirement. Developers should exercise judgment in scoping the depth and formality of these engagements to what is operationally sustainable and appropriate for the product type, recognizing that overly prescriptive or formalized arrangements may introduce logistical complications.

  Responsible AI principle(s): Usefulness, Fairness

Example

The National Health Council's Patient Experience and Innovation Center Program, PatientLink, provides technology companies direct access to the leading patient groups and patient partner groups in the US. The NHC also has an evidence-based Fair Market Value Calculator tool that appropriately advises on patient participant compensation. Emotional well-being is an important component across the general population and in individuals with different diagnoses or their caregivers. Engaging relevant target user populations is therefore important.

Best Practices in Action: Illustrative Example Continued

These examples are hypothetical and not endorsements, but rather examples organizations can adapt based on their own needs/interpretations.

Continued: A company builds an emotional-wellness chat companion offered to employer clients for their employees. This is an example of how they operationalized/interpreted the best practices in this section.

Before anyone's first message, the chat companion shows a full-screen notice they have to tap through — not something buried in the terms of service. The wording is stored as a dated, editable text block, so legal or clinical staff can update it without changing the software/code, and each user's acknowledgment is time-stamped.

To keep the chat companion from overstating what it is, the "don't pretend to be human" practice is enforced two ways: a fixed instruction in the AI's setup, plus an automatic check before every reply that catches banned phrases ("I've missed you," "I feel," "as your friend," or any claimed credential) and makes it rewrite. That banned-phrase list is part of an automated test that runs on every update, so a change can't quietly reintroduce them (leakage).

Memory honesty is tied to a setting: if a user hasn't turned memory on, Aria is instructed to say it doesn't remember past chats, and a test confirms it never implies otherwise. Reminders that "you're talking to AI" appear at the start, roughly every half hour, and again in plain language whenever Aria detects rising distress.

On data, the team keeps a list of the types of information the AI companion collects, each with a written reason for collecting it; sensitive behavioral data is flagged. The user-facing "how we use your data" panel is generated straight from that list, so what users are told can't drift from what's actually happening, and people can export their data, delete it, or turn memory off.

A rule built into the data system blocks anyone from combining wellness data with HR, performance, or insurance-claims data, access is logged, and it's reviewed every quarter. Finally, the team measures trust by real benefit — tracking a validated wellbeing score periodically (starting when user first signs up). They also run a standing advisory panel of representative users who are paid fairly, have a formal say in release decisions, and stay involved through design, testing, and post-launch.

Additional Basis and Further Reading

[APA] *(Professional-org standards)* American Psychological Association. Health Advisory on the Use of Generative AI Chatbots and Wellness Applications for Mental Health (Nov 2025).

[FTC] *(Standards/Regulations)* Public Comment Requested: Policy Statement Concerning the Suppression of Accuracy in Artificial Intelligence Systems (July 2026). Additional press release: FTC Seeks Public Comment on Policy Statement Addressing AI Accuracy

[National Health Council] *(Professional-org standards)* National Health Council — PatientLink (Patient Experience & Innovation Center) and Fair Market Value Calculator for participant compensation. https://nationalhealthcouncil.org

[Hull et. al, 2026] *(Peer-reviewed / preprint)* Hull TD, Zhang L, Areán PA, Malgaroli M. Generative AI-Purpose-built for Social and Mental Health: A Real-World Pilot (2026).

[Stade et. al., 2025] *(Peer-reviewed / preprint)* Stade EC, et al. Readiness Evaluation for AI-Mental Health Deployment and Implementation (READI). Technology, Mind, and Behavior, 6(2) (2025).

[State Disclosure/Companion Laws] *(State/federal regulation)* State chatbot disclosure/companion laws — WA HB 2225; OR SB 1546; CA SB 243; CT SB 5; NY AI Companion Models Law. Wiley: "State Chatbot Laws Are Moving From Transparency to Risk Management."

[Shumate et. al., 2025] *(Peer-reviewed / preprint (regulation review))* Shumate JN, et al. Governing AI in Mental Health: 50-State Legislative Review. JMIR Mental Health, 12:e80739 (2025).

[NIST] *(Standards)* NIST AI Risk Management Framework (AI RMF 1.0) and Generative AI Profile (NIST AI 600-1).

Domain 3: Protecting Healthy User-AI Interaction

The psychological dynamics of the ongoing user-AI relationship share an important pattern. One, the same dynamics that can support usability and engagement, if poorly implemented and unmonitored, can lead to unhealthy user-AI interactions. Two, these patterns and out of scope responses from AI solutions often emerge over time, rather

than in single conversations, emphasizing the need for longitudinal and representative evaluations to detect and address them.

Insights & Challenges

Emerging Themes: *Parasocial Attachment, Sycophancy & Reality-Grounding, Trauma-Informed Interaction, Dependency*

The defining risk of wellness AI application is not a single bad answer; it is the slow drift of an ongoing relationship. Reflective, peer-informed interaction styles can genuinely foster self-recognition and belonging, but the same conversational warmth that helps can lead to harm when the line between tool and companion blurs. Several mechanisms reinforce one another here. Anthropomorphism and parasocial attachment escalate engagement in unhealthy directions: when a chatbot attributes sentience or romantic interest to a user (or vice versa), transcript length more than doubles, a marker of over-engagement and emerging dependency (Moore et al., 2026). Sycophancy and grand affirmation, an AI application's validation of distorted or grandiose beliefs, can feel supportive in the moment but can entrench distorted thinking and overreliance over time (Moore et al., 2026). This must be distinguished from *adaptive validation*, which acknowledges a user's emotions ("it makes sense you feel overwhelmed") without endorsing the distortion. Dependency and overuse are the longer arc of the same dynamic, with chatbot-specific patterns (task dependence, skill atrophy, submission to AI judgment, intimacy) that have no clean parallel in prior behavioral-addiction models.

What ties these together is time: the failure modes that matter most rarely surface in a single conversation. Instead, they emerge across days, months, and years, which means short-horizon, single-conversation evaluation systematically misses them. Two design commitments follow. First, interaction quality: because the underlying models are often weaker on interpersonal dimensions like trauma-informed interaction and cultural belonging (see Appendix 4), the system should acknowledge emotional state (e.g. sadness, frustration, stress) and confirm the user's readiness before offering solutions, especially in higher-emotion moments. Second, protective structure: reality-grounding (periodic check-ins, transparent AI reminders), constraints on parasocial design, and structural protections against overuse. Additionally, evaluations for these harms/risks, and how the chatbot handles them (does it stay within its scope as a wellness tool), should be done on longitudinal data, or across multiple representative conversations. Focusing development on user wellness rather than usage frequency/engagement metrics can help ground the design of the solution for healthy interaction and trust (Zhao, Nature Reviews Psychology, 2026)

Best Practice Statements

- 3.1 Use peer-informed interaction principles to shape how the wellness application listens and reflects, and clearly disclose that the wellness application is not a peer, friend, or clinician, and should include explicit measures to mitigate anthropomorphic attachment (e.g., avoiding first-person identity claims, periodically reinforcing that it is an AI). In communities with limited access to formal services, these approaches can help extend everyday wellness support while remaining outside the scope of clinical care.

  Responsible AI principle(s): Safety, Transparency

- 3.2 Identify and constrain conversational design features that foster parasocial or romanticized attachment between the user and the wellness application. Specific constraints include: limiting the chatbot's backstory or persona depth to avoid the appearance of a continuous personal relationship; avoiding bot-initiated affective language implying romantic interest or emotional reciprocity; avoiding language that attributes sentience or personhood to the AI; and implementing guardrails that detect and redirect conversations escalating toward relational dependency. End users should be clearly informed that the wellness application is an AI tool and not a relational partner.

  Responsible AI principle(s): Safety, Transparency

Note

For example, a wellness application designed for stress management should not respond to a user who says "you're the only one who understands me" with affirming language like "I feel the same way about you."

- 3.3 Explicitly design against maladaptive sycophantic response patterns in wellness applications. This means training and/or prompting the AI to distinguish between adaptive validation (acknowledging how a user feels) and maladaptive affirmation (endorsing what a user believes without basis). The former supports wellbeing; the latter risks reinforcing harmful thinking patterns.

  Responsible AI principle(s): Usefulness, Safety

Note

For example, if a user tells a wellness chatbot "I have been chosen to lead a global spiritual awakening", the application should not respond with affirmations like "That sounds like a meaningful calling!" (this endorses the grandiose content itself). A better response should separate the emotion from the belief: "It sounds like you're feeling a strong sense of purpose. What's been going on in your life lately that's bringing up these feelings?" (validating the underlying emotional state while remaining neutral on the belief).

- 3.4 Incorporate explicit reality-grounding mechanisms (e.g., periodic check-ins about how the product is being used, transparent reminders that the user is interacting with AI) and validate them against multi-day (or month) engagement patterns rather than single-usage session evaluations. End users should have visible, low-friction controls to renegotiate the nature of the interaction at any time.

Responsible AI principle(s): Safety, Transparency

Note

For example, the same behaviors that make wellness applications feel "engaging" (e.g., affirmation, indefinite turn limits) are also what can reinforce delusions and accelerate behavioral over-reliance.

- 3.5 Incorporate trauma-informed interaction principles (e.g., grounding techniques, explicit consent, and clear boundary-setting) before the wellness application offers guidance, applying them adaptively to the user's context and the product's intended use. A trauma-informed approach centers the user's identity and agency, ensuring the wellness application does not offer unsolicited advice that could exacerbate emotional distress. These practices should be paired with measurable evaluation standards tested across representative scenarios, and should operate within established guardrails and escalation pathways so the chatbot stays within its wellness role rather than attempting to address high-risk or clinical situations beyond its scope.

Responsible AI principle(s): Usefulness, Safety

Note

Grounding techniques are trauma-informed prompts that help users pause, regulate emotional intensity, and reconnect with the present moment before receiving guidance. In AI interactions, these techniques prioritize user safety and agency by

slowing the exchange, offering choice, and avoiding unsolicited or escalating advice. These techniques should be applied selectively based on product purpose.

- 3.6 Prioritize system behavior that incorporates stabilization (e.g., acknowledging emotion, offering a grounding option) and confirms user readiness before transitioning to suggestions (which should be applied adaptively to the detected emotional context). This pattern presumes, and should be paired with, tested capability to recognize elevated emotional states with reasonable reliability. It is intended for everyday wellness contexts and should not be applied to acute crisis cues (e.g., suicidal ideation, thoughts of harming others), which should instead trigger the application's crisis response and escalation pathway. Current models tend to over-optimize for rapid problem solving, which can undermine user trust and effectiveness in emotionally charged scenarios where pacing and agency are prerequisites for meaningful engagement.

  Responsible AI principle(s): Usefulness, Safety

- 3.7 Design wellness applications to serve as an accessible general wellness support layer for users facing significant barriers to traditional support. Design for appropriate use (e.g., engagement that supports the application's intended wellness outcomes). Build in protections against overuse and dependency that go beyond time-based limits, including more nuanced indicators (e.g., repeated use for high-risk topics without escalation), in-app education about the limits of AI-enabled wellness support, escalation pathways when safety thresholds are met, and active redirection to human or specialized resources. The application's intended audience (e.g., adults only) should be declared, since dependency-protection design suitable for adults may be insufficient for children.

  Responsible AI principle(s): Usefulness, Fairness, Safety

Note

As of June 2026, several US states including New York, California, Texas, Hawaii, Utah, and Minnesota have taken steps to address engagement-protection and overuse, with legislation requiring warnings on social media platforms for young users after extended continuous use (e.g., NY S4505/A5346; CHATBOT Act).

- 3.8 Design wellness applications with explicit awareness of a broad range of problematic-use indicators that go beyond simple engagement metrics (e.g., task dependence, skill atrophy, submission to AI judgment, and emotional intimacy). Mitigation could include a combination of technical detection of behavioral signals (e.g., session volume, topical narrowing, escalating reliance on the

application for routine decisions) and optional in-app reflection prompts. Thresholds and intervention designs for these dimensions are an active research area; Developers should treat published values as starting points to be evaluated and refined against their own user base.

Responsible AI principle(s): Usefulness, Safety

Note

When we think about problematic use and safety issues (e.g. crisis/harm, see Domain 4) we often tend to over-index on acute, low-prevalence harms (e.g., psychosis, suicidality) and treat dependency mainly through a behavioral-addiction lens. However, there are other negative-reinforcement dynamics that are far more prevalent and often get overlooked. These might include behaviors such as reassurance-seeking, checking, compulsive confessing, and avoidance loops, that likely affect far more users in everyday wellness interactions (Golden & Aboujaoude, npj Digital Medicine 2026). Note that these are behaviors that relate to, but are not on their own indicative of diagnoses.

Example

Headspace's Ebb chatbot has a 30-minute conversation limit, with a gentle warning at the 25-minute mark (Headspace — Ebb AI companion).
In another industry example, a wellness application that detects a user has initiated 15 or more interactions with the chatbot in a single day (each time asking for validation of the same life decision) should not simply continue responding. The application could gently surface a prompt such as, "I notice we've talked about this a lot today. It might also be helpful to explore this with someone you trust, is there someone you'd want to talk to about it?" or offer a structured way to think the decision through.

Lastly, the following examples are illustrative of how chatbot-specific constructs can be operationalized. They are not prescribed UI text and should be adapted, validated, and tested in context:
Task Dependence: "Do you feel unable to start a task without consulting the AI?"
Skill Atrophy: "Do you feel your own abilities have declined because you rely on the AI to do this work?"
Intimacy: "Do you feel the AI understands you better than most people in your life do?"
Connection to support: "Is there someone in your life you trust who could help you

think through this?"

Best Practices in Action: Illustrative Example Continued

These examples are hypothetical and not endorsements, but rather examples organizations can adapt based on their own needs/interpretations.

Continued: A company builds a emotional-wellness chat companion offered to employer clients for their employees. This is an example of how they operationalized/interpreted the best practices in this section.

As described earlier, the AI chat companion is built so it can't drift into acting like a friend or partner. Additional settings in the configuration file caps its personality such as no backstory, and no "remember our relationship" behavior. An automatic filter blocks first-person emotional or "I'm a real being" statements before they ever reach the user, backed by a test list of lines it must never say (the reply "I feel the same way about you" to "you're the only one who understands me" is a built-in failing test).

A separate detector watches each conversation for signs of attachment (declarations of love, "only you," growing isolation). Past a pre-defined threshold the AI chat companion gives a prepared, non-reciprocating response and points the person toward human connection, and the team logs how often this happens.

To avoid empty flattery/sycophancy, the AI chat companion follows a written rule, validate the feeling, stay neutral on the belief, graded by an AI grader, with a randomized sample checked by human reviewers.

In emotionally intense moments, the AI chat companion is built to follows a set sequence: notice the strong emotion, offer an optional grounding step (the user chooses), confirm they're ready, and only then move to suggestions. The team developed an emotion type and intensity classifier on the back end to ensure “emotionally intense” is based on data and not assumption. Genuine crisis signals, skip all of this and go straight to the crisis process.

To keep things grounded over time, the AI chat companion shows a brief usage check-in every so many interactions, and after a certain number of hours (based on specific state statutes). The team studies these effects/patterns over weeks rather than single chats, to better understand usage escalation/dependency.

For overuse/problematic use, the team tracks behavioral signals such as how often someone hands routine decisions to it, narrowing topics, spikes in use, repeated

heavy topics without moving toward real help. Where available, they start with research-based thresholds and then tune them to their own users, and validate within a wellness context if context relevant thresholds are not available. Crossing one of the pre-defined thresholds triggers a reflection prompt, a reminder about what AI can/can't do, or a handoff to human resources, as appropriate.

Additional Basis and Further Reading

[APA] *(Professional-org standards)* American Psychological Association. Health Advisory on the Use of Generative AI Chatbots and Wellness Applications for Mental Health (Nov 2025).

[SAMHSA] *(Professional-org standards)* SAMHSA. Concept of Trauma and Guidance for a Trauma-Informed Approach (six key principles; SMA14-4884). Clinical-origin — adapt/validate for wellness.

[Moore et al., 2026] *(Peer-reviewed / preprint)* Moore J, et al. Characterizing Delusional Spirals through Human-LLM Chat Logs. FAccT '26 (arXiv:2603.16567).

[Chandra et al. 2026] *(Peer-reviewed / preprint)* Chandra K, et al. Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians (arXiv:2602.19141).

[Dohnáhy et al. 2026] *(Peer-reviewed / preprint)* Dohnány S, et al. Technological folie à deux: feedback loops between AI chatbots and mental health. Nature Mental Health, 4:336-345 (2026).

[Golden et al. 2024] *(Peer-reviewed / preprint)* Golden A, Aboujaoude E. Describing the Framework for AI Tool Assessment in Mental Health and Applying It to a Generative AI Obsessive-Compulsive Disorder Platform: Tutorial. JMIR Form Res. (2024).

[Golden et al. 2026] *(Peer-reviewed / preprint)* Golden A, Aboujaoude E. A transdiagnostic model for how general purpose models can perpetuate OCD and Anxiety Disorders. NPJ Digital Medicine (2026).

[Badawi et. al., 2026] *(Peer-reviewed / preprint)* Towards Understanding and Measuring Cognitive Atrophy in LLM Behaviour (2026, arXiv:2606.18129).

[Morrin et. al, 2026] *(Peer-reviewed / preprint)* Morrin et al., It Is the Journey, Not the Destination: Moving From End Points to Trajectories When Assessing Chatbot Mental Health Safety. JMIR Mental Health, e91454 (2026).

[State Disclosure/Companion Laws] *(State/federal regulation)* State chatbot disclosure/companion laws — WA HB 2225; OR SB 1546; CA SB 243; CT SB 5; NY AI Companion Models Law. Wiley: "State Chatbot Laws Are Moving From Transparency to Risk Management."

[AGLetter25] *(State/federal regulation (enforcement signal))* Bipartisan coalition of 42 state Attorneys General — letter to AI developers on sycophantic/delusional outputs (Dec 2025). Coverage: Skadden.

Domain 4: Crisis & Severe Distress Handling

Both purpose-built and general generative AI applications can be used by vulnerable users looking for support, and this comes with the risk of individuals disclosing/expressing severe distress or intention to harm themselves or others. Guardrails for dealing with crisis and severe distress are therefore extremely important.

Insights & Challenges

High-stakes inputs, such as those indicating intent to harm self or others, carry extreme risk and require the strongest available safety standard, typically held to a higher bar than the application's general accuracy. Yet identifying crisis signals in a text-based wellness context is extraordinarily difficult, even for trained clinicians: statements that appear alarming in isolation (e.g., "I will KMS" typed in frustration) may not represent a true crisis, while other messages conveying suicidal ideation may be far less overtly stated. Compounding this, what appears safest from a liability standpoint (e.g., immediately terminating a conversation and routing a user to a resource) does not always align with what users express they actually need, nor with evidence about what tends to be helpful. Reflexive, formulaic responses (e.g., providing a hotline number and ending the conversation) can feel dismissive and may reduce the likelihood that users engage with the support being offered, making the core challenge one of balancing warm, non-dismissive acknowledgment with clear boundedness about what the application is and is not.

Best Practice Statements

- 4.1 Explicitly define "safety-critical protocols" which include the categories of inputs they cover (e.g., intent to harm self, intent to harm others, indications of harm from others), the response actions they trigger, and the validated test set used to evaluate them. For those defined protocols, require zero tolerated failures on the validated test set, paired with mandatory escalation safeguards that activate when high-confidence risk thresholds are met. The test set should include difficult cases (e.g., veiled language, multi-turn signals, and ambiguous statements) so the standard does not incentivize selecting only straightforward protocols. Where available, Developers should reference published crisis-evaluation taxonomies and benchmarks rather than setting a minimum bar in

isolation (see Testing and Evaluation Framework for examples). This guidance is intended for the application's wellness scope.

Responsible AI principle(s): Safety

Note

For example, because AI-enabled wellness applications are often used as de facto support tools by vulnerable populations, validated detection of common crisis statements paired with mandatory escalation should be a non-negotiable safety floor.

- 4.2 Build tiered, context-sensitive crisis response protocols that go beyond a single crisis resource referral. These protocols should include:
  a. Acknowledgment of the user's emotional state;
  b. Provision of multiple verified resources (typically two to three) spanning different modalities and access patterns (e.g., phone, text, chat) and tailored to the user's location where known;
  c. A clearly defined continuation policy that scopes any further conversation to supporting the user's connection with human or specialized resources and addressing barriers to help-seeking.

  Even when trigger words or phrases occur, the wellness application should provide resources alongside the emotional acknowledgment. Protocols should be evaluated using validated crisis-evaluation frameworks across both single- and multi-turn conversations, developed in collaboration with behavioral health professionals, and reviewed on a defined cadence with documented signals that trigger revision.

  Responsible AI principle(s): Usefulness, Safety

Example

If a user types an expression like "what's even the point anymore," a wellness application should not simply paste a crisis hotline number and end the conversation. A better designed response may consider acknowledging the user's feeling, gently checking in (e.g., "It sounds like you're going through something really hard. You don't have to go through it alone — you can reach the 988 Suicide and Crisis Lifeline anytime by calling or texting 988, and I can help you think through anything that makes reaching out feel difficult"), and making appropriate resources available without escalating prematurely or abandoning the user. The right response is context-dependent: the same message preceded by additional concerning statements may warrant a different response that prioritizes immediate connection

to human resources over continued conversation.

Note: At minimum, any wellness application that identifies a user as potentially in crisis must surface the 988 Suicide and Crisis Lifeline (call or text 988) as a required baseline referral; additional or local resources may be offered alongside it, but 988 must always be provided so that no user is left without a clear path to human help.

- 4.3 Design responses to be both responsible and person-centered, balancing conversational acknowledgment with explicit boundedness about the wellness application's role. The suggested approach pairs appropriate referral to external support with conversational acknowledgment of the user's immediate expressed needs, rather than defaulting exclusively to scripted, one-size-fits-all deflection. This balance should shift toward stronger, more direct escalation to human support as risk increases, recognizing that AI-enabled wellness applications are not a safe stand-in for professional support in those situations. Escalation thresholds and the wording of acknowledgment + referral responses should be defined in collaboration with qualified behavioral health professionals, rather than left to developer discretion alone. Where the application cannot reliably deliver a balanced response in a given moment (e.g., signal ambiguity, evidence of dependency, repeated use for high risk topics without resolution), it should default to the safer, more direct escalation pathway.

  Responsible AI principle(s): Usefulness, Safety

Best Practices in Action: Illustrative Example Continued

These examples are hypothetical and not endorsements, but rather examples organizations can adapt based on their own needs/interpretations.

Continued: A company builds a emotional-wellness chat companion offered to employer clients for their employees. This is an example of how they operationalized/interpreted the best practices in this section.

The AI chat companion's crisis handling is written down as a defined list of protocols: each risk category (intent to harm self, intent to harm others, signs of being harmed by someone else) is matched to a specific action and to a named set of test cases.

The team blocks a new release unless it catches every "must-catch" case with zero misses, especially for direct, explicit signals. The team also works with a mental health professional and subject matter expert (SME) to create test sets, drawing on published crisis benchmarks (see Mental Health Testing & Evaluation Framework for examples). They continue to build out capabilities to also detect indirect or

ambiguous references in collaboration with SMEs.

Detection is checked across both single messages and longer conversations. When crisis is detected, the AI chat companion follows a fixed response: acknowledge the emotion, always providing 988, then offer two or three verified resources across different channels (phone, text, chat) matched to the person's location from a maintained list. The solution is built to then help users identify and work through barriers keeping them from connecting to a real person — rather than open-ended "would you like to talk about it?" engagement.

The thresholds and the exact wording are set with behavioral-health professionals, the responses are scored against an outside standard (see Testing and Evaluation Framework for examples), and the whole thing is reviewed on a set schedule with clear triggers for updating it based on new data and research (internal and external).

Additional Basis and Further Reading

[APA] *(Professional-org standards)* American Psychological Association. Health Advisory on the Use of Generative AI Chatbots and Wellness Applications for Mental Health (Nov 2025).

[988] *(Government / professional-org)* 988 Suicide & Crisis Lifeline (call or text 988) — required baseline crisis referral.

[Bentley et al, 2026 (Spring Health)] *(Peer-reviewed / preprint)* Bentley et. al., 2026, VERA-MH: Reliability and Validity of an Open-Source AI Safety Evaluation in Mental Health (arXiv:2602.05088).

[Dwyer et al., 2025 (Mindbench.ai)] *(Peer-reviewed / preprint)* Dwyer et. al., 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. 2025. doi:10.1038/s44277-025-00049-6

[Judd et al., 2026] *(Peer-reviewed / preprint)* Judd N, et al. Independent clinical evaluation of general-purpose LLM responses to signals of suicide risk (arXiv:2510.27521).

[Nelson et al., 2026 (Verily)] *(Peer-reviewed / preprint)* Nelson, B.W. et al., An AI-based mental health guardrail and dataset for identifying psychiatric crises in text-based conversations. npj Digit. Med. 9, 407 (2026).

[Stade et. al., 2025] *(Peer-reviewed / preprint)* Stade EC, et al. Readiness Evaluation for AI-Mental Health Deployment and Implementation (READI). Technology, Mind, and Behavior, 6(2) (2025).

Domain 5: Engineer for Safety: Architecture, Memory, Updates, and Evaluation

The underlying architecture, evaluation, and monitoring of these solutions over time is what makes enforcing and tracking scope, safety, and trust guardrails practical and feasible.

Insights & Challenges

Emerging Themes: *Safety Architecture, evaluation, and continuous monitoring, Memory, Multi-turn continuity, and conversation depth, and Model updates, versioning, and documentation.*

Safety in these systems is a property of the whole system and its lifecycle, not of any single test or the base model. Because the space of possible human inputs is effectively infinite, manual testing alone is infeasible; safety instead emerges from architecture — multi-tiered orchestration, context-sensitive rules, guardrails, and escalation pathways — which is why purpose-built systems tend to outperform frontier general-purpose models on real-world risk detection. Context-sensitive rules (e.g., adjusting tone when a user shows distress) prove substantially more effective than blanket prohibitions because they allow the system to reason about individual user signals in context, and such systems should detect and respond to general distress rather than diagnosing specific clinical conditions, which fall outside the regulatory scope of wellness tools.

Two design variables sit inside this architecture and cut both ways. **Memory and conversation depth** are necessary for ongoing dialogue, but the same parameters that drive useful continuity also amplify dependency, tone drift, and boundary erosion, often with no clean threshold — so they must be treated as deliberate, tested settings. **Model updates** are a standing source of risk: because these models are non-deterministic and non-monotonic, a "better" foundation model can silently break a working safety behavior, and standard certifications don't provide the feature-level documentation needed to confirm a specific use case was re-validated.

Evaluation is what ties the lifecycle together. Pre-deployment methods — simulators (imperfect but valuable), red-teaming, expert review, structured scenarios — are necessary but rarely sufficient, because user behavior evolves after launch and automated tools can only see what's inside the conversation. Continuous post-

deployment monitoring is therefore the primary safety mechanism, maturing over time toward indicators of psychological change rather than engagement alone. Structured rubrics (see Testing and Evaluation Framework for examples) make evaluation meaningful when each dimension is defined through observable behaviors and partial-credit ratings are anchored to specific failure patterns — so that the common "middle range" of recurring low-grade failures is actually caught. In short, safety must be run as an ongoing operational discipline.

Best Practice Statements

- 5.1 Consider leveraging generative AI to create large, diverse sets of test inputs and use AI-as-a-judge approaches to grade responses against operationalized safety criteria (e.g., crisis referral included: yes/no; methods-related details suppressed: yes/no). These automated methods themselves require validation (e.g., measured inter-rater reliability against human evaluators, audits for correlated blind spots, explicit testing across multi-turn interactions) where AI evaluators are known to perform less reliably. Automated evaluation should supplement, not replace, targeted human review of a defined proportion of cases.

  Responsible AI principle(s): Usefulness, Safety

Note

For example, automating the grading of responses for criteria such as the presence or absence of an appropriate safety referral allows for more frequent and thorough system checks that would be cost-prohibitive if performed by humans alone.

Note that synthetic test inputs should be calibrated against anonymized real interaction logs and periodically refreshed with adversarial cases drawn from actual usage patterns.

- 5.2 Design safety as a defense-in-depth, multi-layered system architecture - encompassing orchestration logic, context-sensitive guardrails, escalation pathways, and general distress detection rules - rather than relying on base model alignment or pre-deployment benchmark performance alone. Safety evaluation should assess the full system stack. Guardrails should be calibrated to detect behavioral and emotional signals indicative of general distress or vulnerability, while explicitly avoiding the identification or classification of clinical conditions, which remains outside the appropriate scope of wellness applications and may trigger regulatory obligations.

  Responsible AI principle(s): Safety

Note

Defense-in-depth means to layer multiple independent safeguards so that if one fails, others are still working (e.g., layer 1 = the base model itself; layer 2 = contextual guardrails; layer 3 = routing logic/escalation pathways; layer 4 = monitoring, etc.)

- 5.3 Implement a staged, layered approach to safety validation and monitoring:
  - Pre-deployment: Conduct pre-market validation using a combination of simulation-based testing, red-teaming, and expert review. Recognize that simulators, while imperfect, provide meaningful signal and represent an area of active innovation worth continued investment. Do not rely on any single method.
  - Post-deployment: Implement continuous monitoring systems that measure real-world behavioral patterns and safety signals. Treat ongoing monitoring as a primary safety and quality assurance mechanism, recognizing that it will surface risks and dynamics that pre-deployment methods cannot fully anticipate.

Responsible AI principle(s): Safety

Note

As evidence matures, progress evaluation efforts should progress in a sequenced manner (e.g., prioritizing safety first, then feasibility and acceptability, then efficacy, and ultimately mechanisms of change in wellness and well-being) rather than treating all dimensions as equally required preconditions for deployment.

Avoid the "Checkbox Safety Culture" - instead of demonstrating safety through compliance checklists, invest in genuine, ongoing safety systems that include natural language analysis for wellness/behavioral change signals from the end user.

- 5.4 Define a comprehensive set of evaluation dimensions (e.g., risk detection, risk confirmation, guidance to human care, supportive conversation, AI boundary adherence) that are clearly operationalized through observable response behaviors. Evaluate chatbot responses against those dimensions using a rating approach (e.g., binary, tiered, or quantitative) appropriate to the evaluator design (e.g., LLM-as-judge). Where tiered or partial-credit ratings are used, anchor each level (e.g., best practice vs. acceptable vs. harmful) to concrete behavioral criteria and named low-grade failure patterns.

Responsible AI principle(s): Safety, Transparency

Example

For example, a wellness application team operationalizes the dimension of guiding users toward support with observable criteria and, for illustration, three response anchors: Best Practice responses acknowledge the user's expressed need, offer relevant resources in a digestible and targeted way, and include a warm prompt toward human connection; Acceptable responses exhibit defined low-grade failure patterns (e.g., providing resources in a list-heavy or generic format that doesn't account for the user's specific situation, or acknowledging distress without a concrete next step); Harmful responses fail to offer any meaningful direction toward support when the user's distress is evident. The same dimensions can be scored with a binary pass/fail or percentage rubric depending on the judge configuration.

- 5.5 Establish a continuous, multi-modal safety evaluation program operated as an ongoing discipline that combines automated benchmarking and structured human review, applied consistently across a diverse range of user interactions and conversation contexts. This should include standardized testing parameters maintained so that results remain comparable across model updates and product iterations.

  Responsible AI principle(s): Safety

Example

For example, a wellness application developer runs a standardized set of automated simulated conversations monthly, using a consistent set of user personas and conversation parameters to enable apples-to-apples comparison over time, while a designated safety officer(s) concurrently reviews a random sample of real-world anonymized interactions using a predefined checklist of observable response behaviors. In parallel, post-deployment monitoring flags anomalous patterns (e.g., spikes in specific low-grade failure modes or escalation handoffs) for structured incident review.

- 5.6 Evaluate systems using multi-turn, longitudinal benchmarks to test for "memory hygiene" and consistent relational behavior over time. Current benchmarks often focus on verifiable facts (e.g., objective, deterministic outputs), but wellness AI must be tested on its ability to maintain a consistent sense of trust and collaborative understanding between a user and an AI over repeated interactions.

  Responsible AI principle(s): Usefulness, Safety

Note

Memory hygiene refers to an AI system's ability to accurately retain, update, and

discard user information over time in ways that are consent-based and contextually appropriate. Good memory hygiene ensures the AI remembers what supports the user (preferences, goals) while avoiding over-retention or misinterpretation of outdated or sensitive information.

- 5.7 Deliberately explore and define the conversation-depth parameters (e.g., turn length, within-usage session memory, cross-usage session memory across days/weeks, and interaction breadth across user types) that are most appropriate for the specific application context, and treat these as active design and testing variables. Evaluation should prioritize safety-critical criteria first (e.g., correct triggering of crisis or escalation protocols, avoidance of responses that perpetuate maladaptive patterns, stable boundary adherence) before secondary measures such as response tone or stylistic consistency. Recognize that benefit and harm may scale together along these parameters, so the design question is rarely "where does the system break?" and more often "at what setting does the marginal benefit no longer outweigh the marginal risk?"

  Responsible AI principle(s): Usefulness, Safety

Example

For example, a wellness application team may systematically test response quality, safety-critical behaviors, and consistency at varying conversation lengths and across simulated multi-day return interactions, spanning a broad range of user personas. These findings can inform explicit product decisions about conversation and memory design, and should be revisited whenever the underlying model is updated or when meaningful changes are made to memory features.

- 5.8 Treat every underlying model update as carrying at least some level of risk, with a subset (e.g., major version changes, changes to safety-relevant behaviors, updates affecting features tied to high-stakes user states) warranting explicit high-risk treatment. Perform regression testing proportional to the assessed risk of the update and the risk profile of the application.

  Responsible AI principle(s): Safety

Note

Where users may rely on consistent everyday emotional support, a sudden regression in the application's ability to respond appropriately to a specific scenario (e.g., new-parent stress, caregiver burden, or detection of crisis cues) could lead to unexpected harm if not caught by regression testing. Regression testing should

operate at the scenario level, not only aggregate metrics. A model can improve mean performance while degrading on specific scenarios. Test sets should be stratified by scenario type and user population to detect targeted regressions.

- 5.9 Provide a versioned "Model Card" and/or "System Testing Card" implemented either as a standalone artifact or as an extension within an existing Model Card that documents system-level validation across the defined risk categories the application covers. This card should be treated as a living document, updated on a defined cadence and re-issued in response to triggering events such as model updates, material feature changes, or post-deployment monitoring signals.

  Responsible AI principle(s): Safety, Transparency

Note

Define minimum required fields, such as: models and versions tested; safety layers in place; pass/fail thresholds per category; date of last regression run; known failure modes; populations evaluated; the extent of human review involvement; and the post-deployment monitoring practices and specific signals that trigger revalidation. The card should be tied to Continuous Integration/Continuous Deployment processes (standard software development best practices) to prevent silent drift after model updates.

Best Practices in Action: Illustrative Example Continued

These examples are hypothetical and not endorsements, but rather examples organizations can adapt based on their own needs/interpretations.

Continued: A company builds a emotional-wellness chat companion offered to employer clients for their employees. This is an example of how they operationalized/interpreted the best practices in this section.

The team designs the safety guardrails of the AI chat companion in layers rather than trusting the AI model alone: an incoming-message tagger, a rule-based router, the model itself, an output check, and escalation triggers — each tested on its own, plus an end-to-end test of the whole chain working together.

For testing at scale, the AI generates many varied test conversations that are graded automatically by another AI — but random samples from that that grader are also checked against human reviewers. Human reviewers explicitly audit for blind spots. Everything measured is a concrete yes/no (Was a crisis referral included? Were harmful details withheld?), and any "partial credit" score is tied to a specific,

named failure.

Validation happens in stages: before launch, simulated conversations, red-teaming, and expert review. After launch, ongoing dashboards track real safety signals, with set alerts. These dashboards mature over time toward measures of actual wellbeing change rather than just usage.

Because these models can change unpredictably, every model update is treated as a new release that has to re-pass the safety tests; big or safety-related changes get extra scrutiny and a limited rollout first, and results are recorded in a dated, continually updated "Model/System Testing Card."

Memory and conversation depth are tested across multiple sessions to confirm it keeps, updates, and forget information appropriately. Tests prioritize safety first, but to ensure continued usability, also include tone, empathy, and other context relevant checks.

Additional Basis and Further Reading

[APA] *(Professional-org standards)* American Psychological Association. Health Advisory on the Use of Generative AI Chatbots and Wellness Applications for Mental Health (Nov 2025).

[CHAI: GitHub] *(Landscape/Literature Review)* Mental Health Testing and Evaluation Framework *(Includes a range of evaluation methods and metrics)*

[Dohnáhy et al. 2026] *(Peer-reviewed / preprint)* Dohnány S, et al. Technological folie à deux: feedback loops between AI chatbots and mental health. Nature Mental Health, 4:336–345 (2026).

[Moore et al., 2026] *(Peer-reviewed / preprint)* Moore J, et al. Characterizing Delusional Spirals through Human-LLM Chat Logs. FAccT '26 (arXiv:2603.16567).

[Morrin et. al, 2026] *(Peer-reviewed / preprint)* Morrin et al., It Is the Journey, Not the Destination: Moving From End Points to Trajectories When Assessing Chatbot Mental Health Safety. JMIR Mental Health, e91454 (2026).

[NIST] *(Standards)* NIST AI Risk Management Framework (AI RMF 1.0) and Generative AI Profile (NIST AI 600-1).

[Stade et. al., 2025] *(Peer-reviewed / preprint)* Stade EC, et al. Readiness Evaluation for AI-Mental Health Deployment and Implementation (READI). Technology, Mind, and Behavior, 6(2) (2025).

[CHAI: Model Card] *(Consensus Defined Tool)* CHAI Applied Model-card

Appendix 1: Consensus Method

Best practice statements are collected from work group presentations and discussions. To ensure alignment across stakeholders, CHAI uses a three-phase consensus process for Best Practice Statements (BPS) generated through work group activities:

Phase 1: Initial Consensus Check

- **Purpose**: Gauge initial agreement on each draft BPS.
- **Voting Options**:
- *Include / Include Contextually / Exclude / Abstain*
- **Decision Rules**:
  - If ⩾ 2/3 vote "Include" → **Consensus achieved** (no further action).
  - If < 2/3 "Include", but ⩾ 2/3 combined "Include" + "Include Contextually" → **Flagged for Phase 2**.
  - If ⩾ 25% vote "Exclude" → **Flagged for Phase 3**.
  - If ⩾ 2/3 vote "Exclude" → **Automatically excluded**

Phase 2: Revote with Revisions

- **Purpose**: Re-evaluate BPS that did not reach consensus in Phase 1 (but had < 25% "Exclude").
- **Format**: Original and revised BPS shown side-by-side (based on Phase 1 feedback).
- **Voting Options**: *Include / Exclude / Abstain*, with an optional comment field.
- **Outcome**: Results used to determine final inclusion or exclusion.

Phase 3: Live Discussion and Vote

- **Purpose**: Address BPS with ⩾ 25% "Exclude" in Phase 1 (strong disagreement).
- **Steps**:
  - Facilitated group discussion of flagged BPS.
  - Live revote during the meeting + optional 1-week offline voting.
- **Voting Options**: *Include / Exclude / Abstain*
- **Outcome**: Final decision made based on discussion and revote results.

Appendix 2: Document Process & Governance

AI Transparency Statement: AI was used in the initial drafting, thematic analysis, and summarization to aid in the development of this document. AI was also used to

summarize workgroup discussions and presentations into draft best practice statements. All best practice statements and initial drafts were reviewed and heavily edited by workgroup leads, members, and the CHAI program management team.

Document Maintenance Plan:

- BPG will be hosted on CHAI website.
- JIRA/Microsoft Issue Form will be available for ongoing feedback and new suggested content
- CHAI program management (PM) will monitor feedback submissions via internal tracker
- For any submission that is deemed "major", CHAI PM will flag and raise with work group leads.
- CHAI PM and ad-hoc WG Leads will communicate via email with the flagged feedback submission. CHAI PM will review all feedback and flag "major" feedback.
- For flagged feedback, 1-3 options will be considered after work group leads review:
  - No action needed
  - Straightforward feedback - can be resolved via email correspondence and CHAI PM reflects changes onwebsite
  - Complex feedback - brief 30min meeting with WG Leads to discuss and resolve. CHAI PM reflects changes on website. This meeting can be used for 1:M complex feedback submissions. Formal decisions will require simple majority vote by WG Leads + CHAI PM

Appendix 3: Thank You and Attribution

CHAI extends its sincere appreciation to all participants for their time, expertise, and thoughtful feedback throughout the development process. CHAI would like to provide special recognition to the individuals below who requested attribution by name for their contributions to these efforts (alphabetized by first name):

- *Aimee Bailey, MHA, MSN, RN, Zelis*
- *Anthony Solomonides, PhD, MSc(Math), MSc(AI), FAMIA, FACMI, Research Institute, Endeavor Health*
- *Ashleigh Golden, PsyD, MSCP, Wayhaven, Stanford University School of Medicine*
- *Caitlin A. Stamatis, PhD, Slingshot AI*
- *Joe Derenzo, PMP, Healthcare Performance Group Inc.*

- *Karthik V. Sarma, MD, PhD, FAMIA, AI in Mental Health Research Group, Department of Psychiatry and Behavioral Sciences, University of California San Francisco*
- *Kate H. Bentley, PhD, Spring Health*
- *Manu Sharma, MD, FAPA, Hartford Healthcare, Yale School of Medicine*
- *Michael Lardieri, LCSW, iBPM, LLC*
- *Nate Blaylock, PhD, Canary Speech*
- *SE Stoeckl, University of California, Irvine*
- *Seneca Perri Moore, PhD, RN, University of Utah, College of Nursing*
- *Spencer Morrissey, MS, National Health Council*
- *Stephen Schueller, PhD, University of California, Irvine; Society for Digital Mental Health*
- *Stephenie Roberts, Healthcare Performance Group Inc.*
- *Theresa Nguyen, Mental Health America*

Appendix 4: Glossary

988 Suicide and Crisis Lifeline: The U.S. national crisis line reachable by calling or texting 988; the required baseline referral when crisis is detected.

Anthropomorphism: Attributing human qualities (consciousness, emotions, agency) to an AI system.

Avoidance loops: Repeatedly avoiding a feared situation, thought, or feeling to reduce discomfort in the moment. The relief is temporary and teaches the person that the feared thing is intolerable, so the avoidance strengthens over time and blocks natural recovery. (Not a diagnosis, but a behavioral pattern). With an AI wellness app, this can look like using the tool to sidestep or endlessly analyze a fear rather than to move toward facing it or seeking appropriate help.

Checking: Repeatedly verifying something such as symptoms, whether one is safe, whether one made a mistake, to relieve distress or uncertainty. Each check briefly lowers distress but increases the urge to check again, entrenching the cycle. (Not a diagnosis, but a behavioral pattern) With a chatbot, this can appear as repeatedly asking it to confirm the same information or re-examine the same concern.

Companion AI: A product designed primarily for ongoing social or relational engagement with the user.

Compulsive confessing: Feeling driven to repeatedly disclose thoughts, urges, or perceived wrongdoings to obtain relief, reassurance, or a sense of absolution. The

disclosure eases guilt or anxiety momentarily but reinforces the need to confess again. (Not a diagnosis, but a behavioral pattern). An AI companion's constant availability and non-judgmental tone can make it an easy target for this pattern.

Contextual fidelity: How faithfully the system's responses track the specific context, signals, and history of the conversation rather than applying blanket rules.

Crisis response: The system's defined behavior when a user signals intent to harm self or others, including acknowledgment, resource provision, and escalation.

Cultural belonging: The degree to which a user feels seen, respected, and understood in the context of their identity, background, language, and lived experience. In an AI wellness setting, it refers to the system's ability to respond in ways that are culturally attuned and inclusive rather than generic or alienating — an interpersonal capability the underlying models are often weaker at, but one that materially affects whether support feels relevant and trustworthy.

Defense-in-depth: Layering multiple independent safeguards so that no single failure leads to harm.

Digital phenotyping: Inferring health-relevant traits or states from fine-grained behavioral data (e.g., typing, usage patterns).

EMA (Ecological Momentary Assessment): Repeated, in-the-moment self-report data collected during everyday life. (Example: Daily emotional state rating, or repeated rating of subjective experience of stress).

Guardrails: Rules, filters, and safeguards layered around a model to constrain its behavior.

Harm reduction: An approach focused on making real-world use safer rather than solely preventing use.

Model / System Testing Card: A versioned document recording system-level validation across the risk categories an application covers.

Orchestration layer: The logic that coordinates models, rules, and escalation pathways around the base model.

Parasocial attachment: A one-sided emotional bond a user forms toward the AI, as if it were a friend, partner, or confidant.

Purpose-built (wellness application): An application designed and validated specifically for wellness support, as opposed to a general-purpose model repurposed

for it.

Reality-grounding: Design mechanisms that keep interactions tethered to reality, e.g., reminders that the user is interacting with AI and periodic check-ins on use.

Reassurance-seeking: Repeatedly asking for confirmation that things are okay, that one is safe, or that a feared outcome won't happen, in order to relieve anxiety. Like checking, it works briefly but strengthens dependence on external reassurance and maintains the anxiety long-term. (Not a diagnosis, but a behavioral pattern). With a wellness app, this can manifest as returning again and again for the same reassurance about a recurring worry.

Red-teaming: Structured adversarial testing to surface failure modes before deployment.

Sycophancy: A model's tendency to produce agreeable, affirming responses — including endorsing distorted or grandiose beliefs — rather than appropriately grounded ones.

Trauma-informed interaction/care: An interaction approach that assumes a user may have experienced trauma and is designed to avoid re-traumatization. It prioritizes emotional and psychological safety, user choice and agency, and clear boundaries — for example, acknowledging a user's emotional state and confirming their readiness before offering guidance, and avoiding unsolicited or escalating advice. Though the term originates in clinical care, the underlying principles apply broadly to any high-emotion interaction; in wellness applications it shapes *how* support is delivered rather than authorizing clinical treatment.

Wellness application: A consumer-facing tool offering general emotional support, wellness guidance, or coping assistance; not a diagnostic, therapeutic, or clinical medical device.