# AI Safety Evaluation Questions

This document outlines the AI safety and security questions that benefits consultants and employer benefits leaders should ask when evaluating mental health AI solutions.

## 1. How do you validate that your AI tools are safe for mental health contexts, and which standards do you use?

**Why ask this:** AI for mental health requires high safety standards because of the potential for harm. Some vendors rely only on limited internal metrics to demonstrate their AI tool is safe.

**What a strong answer should reveal:** Vendors should use open-source, clinically grounded safety benchmarks, standards, or certifications, and even results from clinical trials when appropriate. They should also be willing to share evaluation methods and results.

## 2. Do you or any third parties retain or utilize user data for external AI model training? And do you maintain a zero-retention data policy?

**Why ask this:** Sensitive mental health data may be processed by third-party large language models. Buyers need to understand whether that data can be retained, reused, or used to train future models and for how long.

**What a strong answer should reveal:** Vendors should operate in a secure, self-hosted, closed-loop environment where user data is not shared with public models, retained by external vendors, or monetized.

## 3. Is there a defined 24/7/365 human clinician escalation path for high-risk AI interactions, and can you demonstrate how this plays out?

**Why ask this:** Mental health AI tools should recognize acute distress and know when to escalate to human support.

**What a strong answer should reveal:** A strong system should automatically flag explicit or ambiguous high-risk scenarios, such as suicidality. Escalations should route quickly to qualified human clinical support, with clear response-time expectations. Vendors must be able to demonstrate that the escalation path is appropriately and consistently invoked.

## 4. Is your AI designed to support clinicians or replace clinical judgment?

**Why ask this:** Buyers should understand whether the AI makes clinical decisions or gives diagnoses, or whether it supports administrative workflows and enhances care under human oversight.

**What a strong answer should reveal:** Vendors should use a strict human-in-the-loop model. For example, AI may reduce clinical burden by drafting notes or synthesizing information, but it should not replace clinical judgment or make autonomous care decisions.

## 5. Are users clearly informed when they are interacting with AI? And can they choose human-only interactions?

**Why ask this:** Transparency is central to responsible AI. People seeking mental health support should know when AI is part of the experience.

**What a strong answer should reveal:** Vendors should require clear, opt-in consent before users or providers engage with AI-supported features. They should always offer a human-only option.

## 6. What multi-layered safety frameworks and compliance controls govern your AI systems?

**Why ask this:** Standard privacy policies and baseline HIPAA compliance may not be enough to oversee complex generative AI systems.

**What a strong answer should reveal:** Vendors should describe a multi-layered governance model across the company, model, and individual interaction levels. They should also identify applicable compliance standards, such as HIPAA, GDPR, SOC 2 Type II, and HITRUST, and explain how clinicians participate in governance.