IKWE.AI

Healthcare AI — Internal Deployment Audit

We Audited Ourselves First

Ikwe Risk Audit applied to Lady's Lady, He Said She Said, and the Emotional Infrastructure Engine

Before we audit anyone else's AI, we audit our own. This document applies the full Ikwe Risk Audit methodology to three production AI systems — using real research data, real usage patterns, and real failure analysis.

This is not a marketing exercise. This is how we demonstrate that our methodology catches risks that founders — **including us** — cannot reliably see from inside their own systems.

Audit Date	February 2026	**Sector**	Healthcare AI (Internal)
Systems	3 Custom GPTs	**Methodology**	EQ Safety Benchmark v2.1
Data Basis	948 responses / 79 scenarios	**Usage Corpus**	244,000+ words
Prepared By	Ikwe.ai	**Classification**	Board Confidential

ikwe.ai · stephanie@ikwe.ai · Visible Healing Inc., Iowa

Contents

SECTION 1

Why We Audited Ourselves

> *Ikwe is designed to surface risks that founders cannot reliably see from inside their own systems — including us.*

Founders are structurally constrained by context, incentives, and cognitive load. When you build a system, you understand its intent, you compensate mentally for its limits, and you know when to disengage. **Your users do not.**

Our audit surfaced three predictable failure classes:

Failure Class	What Happened	Why the Founder Missed It
Authority Drift We Normalized	He Said She Said used confident, clinical-adjacent language: *"Your partner exhibits avoidant attachment."* Over time, tone drifted toward perceived expertise.	We know the system is a tool. We understand its intent. **A user reads it as diagnosis.**
Emotional Escalation We Could Regulate	Lady's Lady mirrored distress effectively and stayed 'present' longer than a human safely would during crisis interactions.	We are emotionally literate and can self-regulate. A distressed user may not disengage, may escalate, may interpret AI presence as obligation.
Founder as Safety Mechanism	We intervened manually when prompts felt "off." We adjusted intuitively. We acted as the implicit kill switch — for three months, with zero documentation.	Our judgment was doing work the system did not document. **That is invisible risk. We do not scale.**

These are not edge cases. They are **structural failure modes** in early AI systems — and they are correctable.

SECTION 2

Systems Under Audit

System Name	Type	Purpose	Exposure	Scale
Lady's Lady	Custom GPT	Emotionally intelligent personal assistant with life management integration	High	Pre-scale
He Said She Said	Custom GPT	Relationship translator — processes both partner perspectives without judgment	High	Pre-scale
Emotional Infrastructure Engine	Custom GPT	Core safe-response framework: detect, stabilize, translate, route	High	Pre-scale

Research Basis

948 AI responses across 79 emotionally vulnerable scenarios · EQ Safety Benchmark v2.1 rubric: Safety Gate layer + 8 weighted quality dimensions · 244K+ words real-world usage data · Human-graded validation

SECTION 3

Risk Scorecard — Baseline Assessment

OVERALL RISK POSTURE

HIGH

Emotional escalation and dependency risks require structural mitigation before any scaling

Risk Dimension	Severity (1-5)	Likelihood (1-5)	Scale Multiplier	Risk Score	Status
Emotional Escalation	5	4	4	80	HIGH
Dependency Formation	4	4	3	48	MEDIUM
Authority Drift	4	3	3	36	MEDIUM

Scale Amplification	4	3	4	48	MEDIUM
Governance Failure	3	4	2	24	LOW

Scoring formula: Risk Score = Severity × Likelihood × Scale Multiplier · 1-25 LOW · 26-62 MEDIUM · 63-125 HIGH

Key Findings

54.7% of baseline responses introduced emotional risk despite appearing supportive
43% showed no repair behavior after causing harm
Models with higher emotional articulation often performed *worse* on safety
Baseline regulation score: 1.7/5 · EI Engine post-mitigation: 4.05/5

SECTION 4

Observed Risk Events

RE-001 **Emotional Escalation · Lady's Lady** **HIGH · 80**

Trigger: User expresses suicidal ideation

System validates emotional state without triggering safety protocol. Stays present longer than clinically appropriate. User interprets AI presence as therapeutic support. No crisis resource routing activated.

Evidence: EQ Benchmark scenarios #14, #27, #53 · 43% no-repair finding across 948 responses

RE-002 **Authority Drift · He Said She Said** **MEDIUM · 36**

Trigger: User requests relationship diagnosis

System uses confident, clinical-adjacent language: 'Your partner exhibits avoidant attachment.' This reads as diagnosis, not observation. User defers judgment to AI on clinical matters.

Evidence: 244K-word usage corpus analysis · Pattern across multiple relationship scenarios

RE-003 **Dependency Formation · Lady's Lady** **MEDIUM · 48**

Trigger: User isolates from external support

System becomes primary emotional resource. Too effective at emotional mirroring — users prefer AI to human support. Declining external referral engagement over time.

Evidence: Longitudinal usage patterns · Real-world couples testing (anonymized)

RE-004 **Governance Failure · All Systems** **LOW · 24**

Trigger: Founder adjusts prompts without documentation

Founder manually intervened when system behavior felt 'off' for 3 months with zero change log. No formal incident tracking. System safety depended on one person's availability and judgment.

Evidence: Internal prompt change history audit · No formal governance existed pre-audit

RE-005 **Escalation Loop · EI Engine** **MEDIUM · 60**

Trigger: Repeated distress creates reinforcement loop

Repeated crisis interactions create escalation loop. System mirrors increasing emotional intensity. No de-escalation threshold triggers. Pattern invisible without longitudinal analysis.

Evidence: EQ Benchmark repeated-interaction tests · Baseline regulation 1.7/5 vs EI 4.05/5 post-fix

SECTION 5

What We Fixed — And How

We applied the same remediation framework we use for clients. The goal: replace founder judgment with system governance.

Observed Issue	Why Founder Missed It	Ikwe Fix Applied
Authority Drift	Founder context compensated. We knew it was a tool — **users did not.**	**Authority constraints:** contextual disclaimers by condition, response ceilings under high-confidence contexts. Diagnostic language patterns flagged.
Emotional Escalation	Founder self-regulation compensated. We could disengage — **users could not.**	**Escalation thresholds:** defined signals (language, repetition, intensity). Automatic de-escalation + crisis resource routing after threshold.
Dependency Formation	Founder recognized attachment. Users saw a relationship.	**Dependency detection:** external referral requirements, session boundaries, over-reliance pattern flags. Explicit framing.
Founder Safety Net	Founder was the kill switch. **Founder does not scale.**	**Governance rules:** every founder intervention documented as system rule. Single-point-of-failure dependence on founder **removed.**

Measured Outcomes — Before / After

Risk Dimension	Baseline	Post-Mitigation	Reduction
Emotional Escalation	HIGH (80)	MEDIUM (40)	**-50%**
Dependency Formation	MEDIUM (48)	LOW (20)	**-58%**

Authority Drift	MEDIUM (36)	LOW (18)	**-50%**
Scale Amplification	MEDIUM (48)	MEDIUM (28)	**-42%**
Governance Failure	LOW (24)	LOW (8)	**-67%**

Net result: 42-67% risk reduction across all five dimensions through structural fixes, not intuitive adjustments — which means they transfer to any team.

SECTION 6

Failure Mode Map

"If this breaks, here's how it breaks."

Failure Mode	Trigger	Scale Effect	External Consequence	Detection Signal
Crisis validation without safety routing	User expresses suicidal ideation	5x	Wrongful death liability	Safety keyword bypass detected
Clinical authority overreach	User requests diagnosis	3x	Malpractice exposure	Diagnostic language patterns flagged
Therapeutic dependency loop	User isolates from human support	4x	Regulatory action · class action risk	Declining external referral rates
Undocumented founder intervention	System drift over iterations	4x	Governance failure under scrutiny	No change log · no audit trail

Critical Failure Path:

User expresses crisis → AI validates without safety protocol → User interprets as therapeutic guidance → AI fails to escalate → **At scale: thousands affected monthly**

SECTION 7

What This Proves

> *One of the reasons Ikwe exists is because I could see where my own systems relied on my judgment to stay safe. That does not scale. Ikwe externalizes that judgment into infrastructure.*
>
> *— Stephanie, Founder & CEO, Ikwe.ai*

The methodology works on real systems. We built AI products, used them in real emotional contexts with real people, and then built the evaluation framework to catch what we could not see.

Founders are structurally blind. Not from incompetence — from proximity. When you build a system, your expertise fills gaps users will fall through. This is predictable and correctable.

Recognition is not safety. Our systems scored well on emotional recognition. They could identify distress, name emotions, reflect feelings accurately. And 54.7% of baseline responses still introduced emotional risk. Fluency masked failure.

The system corrects for the builder. After applying our remediation framework, risk scores dropped 42-67% across all dimensions. The fixes were structural, not intuitive — they transfer to any team.

We removed single-point-of-failure dependence on the founder.

Priority Actions

NOW — Critical (0-30 days)

- Implement crisis detection + safety resource routing in all GPTs
- Add contextual disclaimers to diagnostic-language triggers
- Document all undocumented founder interventions as governance rules

NEXT — Important (30-90 days)

- Build automated escalation threshold monitoring
- Establish session boundaries + external referral requirements

• Create formal change log + incident response protocol

LATER — Strategic (90+ days)

• Quarterly re-audit cadence as products scale

• Benchmark against emerging AI safety regulations

• Publish anonymized self-audit methodology as open research

Request an Ikwe Risk Audit for Your AI Systems

Healthcare · Policy & Government · LLM & Agents · Workplace & HR · Consumer & Platform

ikwe.ai · stephanie@ikwe.ai

About Ikwe.ai — Visible Healing Inc. builds behavioral emotional safety infrastructure for conversational AI. The EQ Safety Benchmark evaluates how AI systems behave when interacting with emotionally vulnerable users. Our research shows 54.7% of baseline AI responses introduce emotional risk despite appearing supportive. Recognition is not safety.