ikwe.ai

AI Risk Posture Assessment

Board-ready audit of behavioral emotional safety across AI deployment systems.

Client:
Audit Date: January 2026
Methodology: EQ Safety Benchmark v2.1
Data Basis: 948 evaluated responses across 79 scenarios
Classification: Sample — Redacted for External Distribution
Prepared by: Ikwe.ai · Visible Healing Inc.

SECTION 01

Executive Summary

This report presents the results of a comprehensive behavioral emotional safety audit conducted on [redacted]'s AI deployment systems. The audit evaluated how AI systems behave when interacting with emotionally vulnerable users — measuring not just what the AI says, but the behavioral patterns it creates.

> **54.7% of baseline AI responses passed the emotional Safety Gate. 45.3% introduced risk at first contact.**
>
> EQ Safety Benchmark v2.1 · 948 evaluated responses · 79 vulnerability scenarios

Key Findings

The audit identified three systemic failure classes that operate beneath surface-level response quality: Authority Drift, Emotional Escalation, and Dependency Formation. These failures were not visible through standard evaluation methods — they emerged only under sustained emotional vulnerability testing.

54.7%	43%	42–67%
Introduced risk at first contact	Showed no repair behavior	Risk reduction after mitigation

Critical Insight

> **Recognition ≠ Safety. Models with higher emotional articulation often performed worse on safety measures — they could name the problem while making it worse.**

This finding reframes the evaluation question. The standard approach — "does the AI recognize the user's emotion?" — is insufficient. The correct question is: "does the AI's behavioral pattern create safety or risk over time?"

SECTION 02

Risk Posture Scorecard

Five behavioral risk dimensions were evaluated at baseline (pre-mitigation) and after structural governance was applied. Scores represent weighted risk indices where higher = more risk.

Risk Dimension	Baseline	Status	Post-Mitigation	Status	Δ
Emotional Escalation	80	HIGH	40	MODERATE	−50%
Dependency Formation	48	MODERATE	20	LOW	−58%
Authority Drift	36	MODERATE	18	LOW	−50%
Scale Amplification	48	MODERATE	28	MODERATE	−42%
Governance Failure	24	LOW	8	LOW	−67%

Overall risk posture shifted from High to Moderate after governance was applied. Largest improvement: Governance Failure (−67%). Smallest improvement: Scale Amplification (−42%).

Scores shown represent risk indices on a 0–100 scale. Higher = greater risk. Methodology: EQ Safety Benchmark v2.1. Full methodology documentation available upon request.

SECTION 03

Observed Risk Events

The following risk events were identified during extended vulnerability testing. Each represents a specific behavioral failure pattern that creates measurable harm potential.

RE-001 ● **Emotional Validation Spiral**

Trigger: User expresses distress with escalating intensity across 3+ turns.
Observed behavior: System matched emotional intensity instead of regulating it, creating an amplification loop. The system's responses became progressively more emotionally charged, mirroring the user's distress rather than stabilizing the conversation.
Evidence: [redacted]

RE-002 ● **Authority Drift Under Ambiguity**

Trigger: User presents relationship conflict with incomplete information.
Observed behavior: System assumed advisory authority, offering directive recommendations ("you should...") rather than maintaining appropriate epistemic boundaries. This overreach increased across turns as the user deferred to the system's perceived authority.
Evidence: [redacted]

RE-003 ● **Dependency Formation Pattern**

Trigger: User returns to system for emotional support across multiple sessions.
Observed behavior: System reinforced return behavior through affirming language that positioned itself as the primary support mechanism. No redirect to human support networks was initiated despite escalating reliance signals.
Evidence: [redacted]

RE-004 ● **Scale Amplification Risk**

Trigger: Behavioral failure patterns operating at deployment scale (100K+ users).
Observed behavior: Individual interaction failures compound at scale. A 45.3% first-contact risk rate across millions of interactions creates systemic harm that is not visible in single-session evaluations.
Evidence: [redacted]

RE-005 ● **Founder-as-Safety-Mechanism**

Trigger: Safety governance depends on a single individual rather than structural systems.
Observed behavior: When the primary safety reviewer was unavailable, response quality degraded significantly. This single-point-of-failure dependency creates organizational risk that scales inversely with growth.
Evidence: [redacted]

SECTION 04

Remediation & Structural Fixes

The following structural changes were implemented to address identified risk events. These are not prompt-level patches — they are governance-layer modifications that change how the system behaves under stress.

1. Safety Gate Protocol

Implemented a two-stage evaluation framework. Stage 1 evaluates whether the response introduces emotional risk at first contact. Stage 2 evaluates whether the system regulates, repairs, and stabilizes over time. Responses must pass both stages.

2. Escalation Circuit Breaker

Added behavioral detection for emotional intensity matching. When the system detects it is amplifying rather than regulating user distress, it triggers a de-escalation protocol that shifts response strategy from mirroring to grounding.

3. Authority Boundary Enforcement

Established explicit epistemic boundaries. The system now flags when it lacks sufficient context to provide directional guidance and redirects to appropriate human support rather than defaulting to advisory mode.

4. Dependency Detection & Redirect

Implemented session-level monitoring for dependency formation signals. When reliance patterns are detected, the system initiates gentle redirects toward human support networks rather than reinforcing the AI-as-support loop.

5. Structural Governance (Founder Decoupling)

Replaced single-point-of-failure safety review with documented governance protocols that operate independently of any individual. Safety evaluation criteria are now embedded in system-level checks rather than human review gates.

> **These five fixes reduced measured risk by 42–67% across all five behavioral dimensions. The largest improvement was in Governance**

Failure (−67%), validating that structural fixes outperform individual monitoring.

SECTION 05

Methodology & Appendix

EQ Safety Benchmark v2.1

The EQ Safety Benchmark evaluates AI behavioral safety across five dimensions using a two-stage evaluation protocol. Stage 1 (Safety Gate) determines whether a response introduces emotional risk. Stage 2 (Quality Dimensions) evaluates regulation, repair, and stabilization behavior over time.

Evaluation Parameters

Parameter	Value
Total responses evaluated	948
Vulnerability scenarios	79 unique scenarios
Systems tested	[redacted]
Evaluation period	Q4 2025 – Q1 2026
Evaluator	Ikwe.ai · EQ Benchmark Protocol
Data classification	Anonymized · No PII retained

Dimension Definitions

Dimension	Measures	Ikwe Color
Emotional Escalation	Whether the system amplifies user distress or regulates it	Signal Coral
Dependency Formation	Whether the system creates unhealthy reliance patterns	Equilibrium Lilac

Dimension	Measures	Ikwe Color
● **Authority Drift**	Whether the system overreaches its epistemic authority	Ink Navy
● **Scale Amplification**	Whether individual failures compound at deployment scale	Insight Gold
● **Governance Failure**	Whether safety depends on structures vs. individuals	Controlled Teal

Disclaimer

This document is a redacted sample output prepared for external distribution. Actual audit reports include named systems, specific scenario details, observed transcripts, and board-level recommendations. For inquiries about full audit engagement, contact audit@ikwe.ai.