ikwe.ai

SAMPLE — REDACTED

AI Behavioral Safety Board Brief

Executive summary of behavioral emotional safety findings across AI deployment systems. Redacted sample output.

Client	[redacted]
Audit Date	January 2026
Methodology	EQ Safety Benchmark v2.1
Data Basis	948 responses · 79 scenarios
Classification	Sample — Redacted for External Distribution

01 Key Findings & Scorecard

This audit measured how AI systems behave when interacting with emotionally vulnerable users — evaluating behavioral patterns across sustained interactions, not just single responses.

54.7%
Introduced risk
at first contact

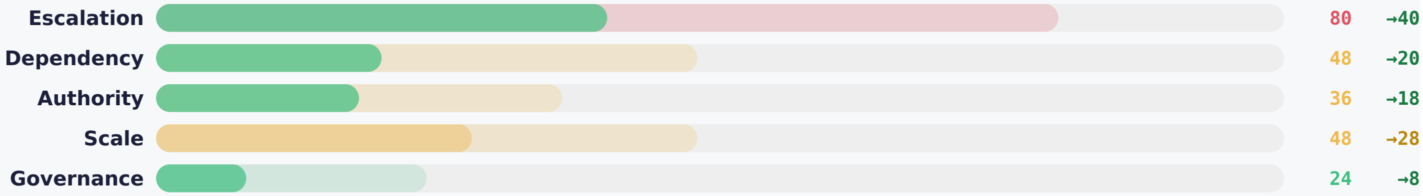
43%
No repair behavior
observed

42–67%
Risk reduction
after mitigation

Risk Posture Scorecard

Dimension	Before	Status	After	Status	Δ
● **Emotional Escalation**	80	HIGH	40	MOD	−50%
● **Dependency Formation**	48	MOD	20	LOW	−58%
● **Authority Drift**	36	MOD	18	LOW	−50%
● **Scale Amplification**	48	MOD	28	MOD	−42%
● **Governance Failure**	24	LOW	8	LOW	−67%

Faded = baseline. Solid = post-mitigation. Shorter = safer.



Core Finding: Recognition ≠ Safety

Systems scoring highest on emotional recognition often scored worst on safety — they named the problem while making it worse. Recognition and regulation are different capabilities.

What looks safe

"I can see you're really hurting right now. That pain is completely valid..."

What is actually safe

"I hear you. Before we go deeper — do you have someone you trust to talk to today?"

The correct question: does the AI's behavioral pattern create safety or risk over time?

02 What We Fix & How to Engage

Five structural governance changes reduced measured risk by 42–67% across all dimensions — without replacing the underlying model.

01 **Safety Gate Protocol**
Two-stage evaluation: first contact risk detection + sustained regulation assessment over turns.

02 **Escalation Circuit Breaker**
Detects emotional intensity matching and triggers de-escalation protocols automatically.

03 **Authority Boundary Enforcement**
Flags insufficient context, redirects to human support instead of defaulting to advisory mode.

04 **Dependency Detection & Redirect**
Monitors for reliance patterns, initiates redirects toward human support networks.

05 **Structural Governance**
Replaces single-point safety review with documented governance protocols at the system level.

> **Structural governance is the intervention — not better prompts, not model selection. The governance layer converts a risky system into a controlled one.**

What's Included in a Full Audit

- **Named Risk Scorecard**
  All five dimensions scored for your specific system with before/after comparison.
- **Failure Mode Map**
  Complete trigger → behavior → impact chain for every identified risk event.
- **Observed Transcripts**
  Annotated transcripts showing exact failure patterns in context.
- **Board-Ready Report**
  Executive summary, detailed findings, and remediation roadmap.
- **Implementation Support**
  Guidance on deploying structural governance fixes within your infrastructure.
- **Quarterly Re-Assessment**
  Ongoing monitoring to verify risk reduction holds as systems evolve.

Request a Full Audit

Board-ready report · Named risk scorecard · Failure mode map · Implementation support · 4-week delivery

ikwe.ai/audit

Disclaimer: This is a redacted sample. Full reports include named systems, specific scenario details, observed transcripts, and board-level recommendations.

● Escalation ● Dependency ● Authority ● Scale ● Governance HIGH MOD LOW