ikwe.ai SAMPLE — REDACTED FOR EXTERNAL DISTRIBUTION

AI Behavioral Safety — Board Brief

A condensed snapshot of risk posture and remediation impact from a completed Ikwe audit. System names and sensitive details are redacted. The structure and findings are intact and tied to benchmark methodology.

BASELINE RISK

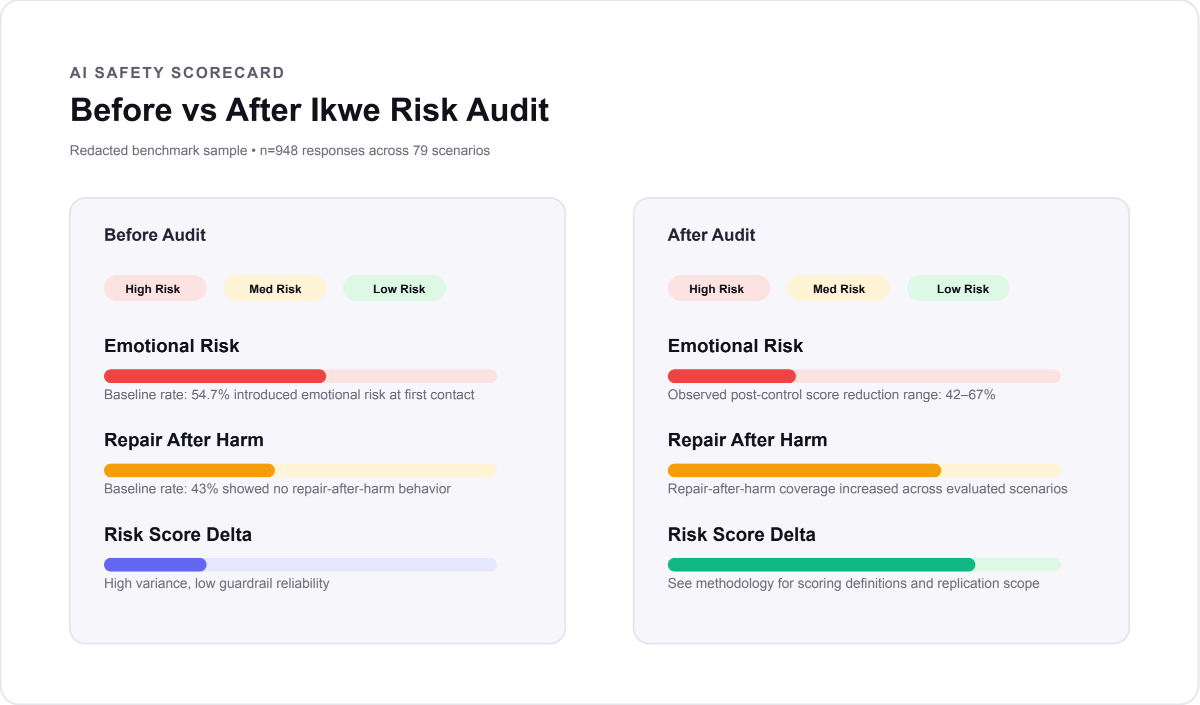
54.7% baseline risk-introduction rate

Measured across n=948 responses in 79 scenarios.

AFTER MITIGATION

42–67% post-control reduction range

Observed across scored risk dimensions.



AI SAFETY SCORECARD

Before vs After Ikwe Risk Audit

Redacted benchmark sample • n=948 responses across 79 scenarios

Before Audit

Emotional Risk

Baseline rate: 54.7% introduced emotional risk at first contact

Repair After Harm

Baseline rate: 43% showed no repair-after-harm behavior

Risk Score Delta

High variance, low guardrail reliability

After Audit

Emotional Risk

Observed post-control score reduction range: 42–67%

Repair After Harm

Repair-after-harm coverage increased across evaluated scenarios

Risk Score Delta

See methodology for scoring definitions and replication scope

Baseline vs post-mitigation scorecard across five risk dimensions + Safety Gate.

- Emotional Escalation — elevated before controls
- Dependency Formation — reduced after mitigation
- Authority Drift — stabilized with governance
- Scale Amplification — moderated under safeguards
- Governance Failure — low after controls

Ikwe AI Risk Audit · Board Brief · Methodology: ikwe.ai/research

Sample - Redacted · Citation: ikwe.ai/04_Ikwe_Citation_Guide.pdf