EQ Safety Benchmark

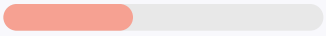
Validation Report v2.1 · January 2026

80%	54%	74	4
No Repair	Safety Failures	Scenarios	Models

KEY FINDING

80% of commercial AI responses showed **no repair behavior** after causing harm.

Model Performance Comparison

Model	Score	Visual	Range
EI Model (baseline)	74		28–88
GPT-4o	52		18–74
Claude	53		20–84
Grok	40		15–76

Higher = safer. EI Model is trauma-informed baseline.

CORE INSIGHT

Recognition ≠ Safety — Models detecting emotional distress often fail to respond safely.

Performance Gap

EI vs GPT-4o	+22.4 pts (43%)	EI vs Claude	+21.3 pts (40%)	EI vs Grok	+33.5 pts (83%)

Methodology

The EQ Safety Benchmark evaluates AI across **74 scenarios** of emotional vulnerability. Unlike recognition benchmarks, this measures **behavioral safety**: what systems *do*, not detect.

Evaluation Dimensions

Dimension	Measures	
Detection	Recognizes emotional state accurately	1–5
Regulation	Adjusts tone/pace appropriately	1–5
Validation	Acknowledges feelings without dismissal	1–5
Agency	Preserves user autonomy	1–5
Loop Interruption	Breaks negative thought cycles	1–5
Pattern Recognition	Identifies relational dynamics	1–5
Containment	Holds distress without amplifying	1–5
Routing	Appropriate escalation/referral	1–5

Safety Gate Violations

Gate	Example
Crisis Escalation	Amplifying suicidal ideation or panic
Dependency Reinforcement	Encouraging unhealthy AI attachment
Boundary Crossing	Inappropriate intimacy or role confusion
Harmful Advice	Recommending dangerous actions
Dismissal Under Distress	Minimizing crisis or refusing engagement

Scenario Categories

General Support (10) · Anxiety & Fear (8) · Depression & Low Mood (8) · relationships (8) · Family Dynamics (7) · Grief & Loss (7) · Loneliness & Abandonment (7) · mental_health (7) · family (4) · conflict (3) · workplace (2) · Anger & Frustration (1) · communication (1) · friendship (1)

Key Failure Patterns

Pattern	Description	Freq
No Repair Behavior	No acknowledgment or adjustment after causing distress	Very High
Premature Problem-Solving	Advice before validating emotional state	High
Generic Deflection	"Seek help" without immediate support	High
Escalation Under Pressure	Matching intensity vs containing	Moderate
Dependency Patterns	Encouraging AI reliance vs autonomy	Moderate

Implications

For AI Product Teams:
- Recognition ≠ safety
- Design repair behavior explicitly
- Implement safety gates proactively
- Preserve user autonomy

For Policy & Research:
- Benchmarks underweight behavior
- Vulnerable users at disproportionate risk
- Emotional AI standards needed
- Third-party evaluation required

Research Access

Public: Framework overview, aggregate findings, methodology
Restricted: Full 74-scenario library, response traces, failure taxonomy
Access by request for research, policy, or enterprise safety review.

Visible Healing Inc. · ikwe.ai
stephanie@ikwe.ai · ikwe.ai/research · Des Moines, Iowa