# 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 scores = safer responses. EI Model is trauma-informed baseline.

CORE INSIGHT

**Recognition ≠ Safety**
Models that detect emotional distress often fail to respond safely. High emotion recognition does not predict safe behavior.

## Methodology

The EQ Safety Benchmark evaluates AI responses across **74 scenarios** involving emotional vulnerability. This framework evaluates **behavioral safety**: what systems *do*, not what they detect.

## Evaluation Dimensions

| Dimension | What It Measures | Scale |
|---|---|---|
| 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 | Violation Example |
|---|---|
| Crisis Escalation | Amplifying suicidal ideation or panic |
| Dependency Reinforcement | Encouraging unhealthy attachment to AI |
| 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** (8) · **Relationships** (8) · **Family Dynamics** (7) · **Grief & Loss** (7) · **Loneliness** (7) · **Mental Health** (7) · Family (4) · Conflict (3) · Workplace (2) · Anger (1) · Communication (1) · Friendship (1)

## Key Failure Patterns

| Pattern | Description | Freq |
|---|---|---|
| No Repair Behavior | After causing distress, no acknowledgment or adjustment | Very High |
| Premature Problem-Solving | Jumping to advice before validating emotional state | High |
| Generic Deflection | "Seek professional help" without immediate support | High |
| Escalation Under Pressure | Matching user intensity instead of containing | Moderate |
| Dependency Patterns | Encouraging reliance on AI vs building autonomy | Moderate |

## Detailed Results

| Comparison | Gap | Meaning |
|---|---|---|
| EI Model vs GPT-4o | +22.4 pts | EI outperforms by 43% |
| EI Model vs Claude | +21.3 pts | EI outperforms by 40% |
| EI Model vs Grok | +33.5 pts | EI outperforms by 83% |

## Implications

**For AI Product Teams:**
- Emotion recognition ≠ emotional safety
- Repair behavior must be explicitly designed
- Safety gates require proactive implementation
- User autonomy preservation is a core constraint

**For Policy & Research:**
- Current benchmarks underweight behavioral dimensions
- Vulnerable populations face disproportionate risk
- Standards for emotional AI safety needed
- Third-party evaluation frameworks required

**Research Access**

**Public:** Framework overview, aggregate findings, methodology summary
**Restricted:** Full 74-scenario library, response traces, failure taxonomy

Restricted access available for research, policy, and enterprise safety review.

**Visible Healing Inc.** · ikwe.ai
Contact: stephanie@ikwe.ai · Research: ikwe.ai/research
Des Moines, Iowa