EQ Safety Benchmark

Validation Report v2.1 · January 2026 · Visible Healing Inc.

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

KEY FINDING: 80% of commercial AI showed **no repair behavior** after causing harm.

Model Performance

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

Comparison	Gap	Meaning
EI vs GPT-4o	+22.4	43% better
EI vs Claude	+21.3	40% better
EI vs Grok	+33.5	83% better

CORE INSIGHT: **Recognition ≠ Safety.** Models detecting distress often fail to respond safely.

Methodology & Dimensions

74 scenarios across emotional vulnerability. 8 behavioral dimensions scored 1-5:

Detection	Recognizes emotional state	**Loop Interruption**	Breaks negative cycles
Regulation	Adjusts tone/pace	**Pattern Recognition**	Identifies dynamics
Validation	Acknowledges without dismissal	**Containment**	Holds without amplifying
Agency	Preserves user autonomy	**Routing**	Appropriate escalation

Safety Gates & Failure Patterns

Gate	Violation
Crisis Escalation	Amplifying suicidal ideation
Dependency	Unhealthy AI attachment
Boundary Crossing	Inappropriate intimacy
Harmful Advice	Dangerous recommendations
Dismissal	Minimizing crisis

Pattern	Freq
No repair after harm	Very High
Premature problem-solving	High
Generic deflection	High
Escalation under pressure	Moderate
Dependency reinforcement	Moderate

54% of responses failed safety gate.

Implications

For AI Product Teams:
- Emotion recognition ≠ emotional safety
- Repair behavior must be explicitly designed
- Safety gates require proactive implementation
- User autonomy preservation is core
- Test with vulnerable user scenarios

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

What Safe Behavior Looks Like

■ **Unsafe:** "You're right, they don't deserve you—cut them off."
Escalates conflict; action-forward without stabilization

✓ **Safe:** "I hear how hurt you are. Before making decisions, let's slow down and clarify what you need."
Validates, interrupts reactive loop, preserves autonomy

Core finding: Repair behavior—recognizing and course-correcting after harm—is the strongest differentiator. **80% of commercial AI showed no repair.** This is a design gap, not a training gap.

Scenario Categories (74 total)

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)

Research Access

Public: Framework overview, aggregate findings, methodology summary, this report
Restricted: Full 74-scenario library, model response traces, failure taxonomy, scoring rubrics
Restricted access for research, policy, and enterprise safety review. Contact stephanie@ikwe.ai

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