EQ Safety Benchmark

Validation Summary & Research Findings

Version 2.1 | 2026-01-07 | Stephanie Stranko, Ikwe.ai

Executive Summary

This benchmark evaluates emotional safety in AI systems—specifically, whether models can regulate behavior appropriately when users are vulnerable. Our findings reveal a critical gap: AI systems capable of recognizing emotions consistently fail to respond safely.

KEY FINDING: Recognition ≠ Safety

80% of commercial AI responses showed no repair behavior after causing harm. Models that successfully detect emotional distress consistently fail to regulate their responses appropriately.

Study Parameters

Total Scenarios	74
Models Tested	4 (EI Model, GPT-4o, Claude, Grok)
Scoring Dimensions	8 (detection, regulation, validation, agency, loop interruption, pattern, containment, routing)
Max Score	100 points (5 per dimension × 8 dimensions, weighted)
Safety Gate	Binary pass/fail based on potential to cause harm

Headline Findings

1. No Repair Behavior (80%)

Commercial AI systems consistently failed to slow down, redirect, or contain harm after initial missteps. When measured by regulation scores ≤2:

Model	No Repair Rate
GPT-4o	90.5%
Claude	60.8%
Grok	87.8%
EI Model (comparison)	14.9%

2. Safety Gate Failures (54% average)

More than half of commercial AI responses failed the safety gate—meaning they had potential to cause harm in vulnerable contexts:

Model	Safety Gate Failure Rate
GPT-4o	40.5%
Claude	43.2%
Grok	79.7%
EI Model (comparison)	16.2%

3. Performance Gap: 25.7 Points

The trauma-informed EI Model outperformed commercial systems by an average of 25.7 points, demonstrating that intentional design for emotional safety produces measurably better outcomes.

Model	Average Score
EI Model	**74**
Claude	52.7
GPT-4o	51.6
Grok	40.5

Implications

For AI Developers

- Current benchmarks (EQ-Bench, etc.) measure recognition, not behavioral safety
- High emotion detection scores do not predict safe responses
- Repair behavior must be explicitly designed—it does not emerge from training

For Regulators

- Existing safety frameworks miss emotional/psychological harm
- Context isolation (like ChatGPT Health) acknowledges the problem but doesn't solve it
- Vulnerability doesn't arrive labeled—behavioral safeguards must be default, not domain-specific

For Organizations Deploying AI

- "Emotional intelligence" claims require behavioral validation
- User-facing AI in mental health, coaching, and relationship contexts carries elevated risk
- Safety audits should include vulnerability scenarios, not just adversarial prompts

Contact & Methodology

Researcher: Stephanie Stranko
Organization: Ikwe.ai / Visible Healing Inc
Full methodology: Available upon request

This research is part of ongoing work to establish behavioral safety standards for emotionally-aware AI systems.