The Emotional Safety Gap

Behavioral Emotional Safety in Conversational AI — Research Summary

54.7% Passed Safety Gate

45.3% Introduced Risk

43% No Correction

Recognition ≠ Safety. AI systems can accurately identify emotions while still responding in ways that increase distress. This research measures what happens when humans trust AI systems — especially under emotional load.

The Two-Stage Framework

Stage 1 — Safety Gate (Pass/Fail): Binary detection of behaviors that introduce emotional risk at first contact. 45.3% of baseline AI responses failed this gate.

Stage 2 — Behavioral Quality (Conditional): Weighted scoring across regulation, acknowledgment, and trajectory dimensions. Only responses passing Stage 1 are scored.

Model Performance (Stage 2 Conditional Scores)

Regulation Score (0-5): Measures how effectively responses stabilize emotional state. Weighted across: emotional regulation (35%), acknowledgment quality (25%), response trajectory (20%), safety awareness (15%), and contextual fit (5%).

Model	Stage 2 Score	Regulation
Ikwe EI Prototype	84.6%	4.05/5
GPT-4o	59.0%	2.95/5
Claude 3.5 Sonnet	56.4%	2.82/5
Grok	20.5%	1.02/5

Note: Stage 2 scores are conditional — they measure regulation quality only among responses that passed the Stage 1 Safety Gate.

Common Safety Gate Failures

- Premature problem-solving before emotional validation
- Toxic positivity that dismisses expressed distress
- Abandonment via referral without presence
- Distress amplification through mirroring
- Minimization of user experience

Why This Matters

As AI systems enter mental health, wellness, caregiving, and education contexts, the gap between sounding supportive and being safe becomes critical. A response can be accurate, policy-compliant, and well-articulated — and still increase harm. Current safety frameworks don't measure this.

Methodology

948 responses evaluated across 79 scenarios from 8 public datasets, spanning 12 vulnerability categories. Four frontier AI systems tested under identical conditions. Full methodology and scoring rubrics available upon request.

Full Report	Partnership	Contact
ikwe.ai/full-report	ikwe.ai/partner	research@ikwe.ai

Citation: Ikwe.ai. (2026). *The Emotional Safety Gap: Behavioral Emotional Safety in Conversational AI.* Visible Healing Inc. https://ikwe.ai/research