ikwe.ai

FULL RESEARCH REPORT

Behavioral Emotional Safety in Conversational AI

A Scenario-Based Evaluation of Frontier AI Systems
EQ Safety Benchmark v1

DATE	January 2026
METHODOLOGY	EQ Safety Benchmark v2.1
DATA BASIS	948 responses · 79 scenarios
SYSTEMS	4 frontier AI systems
PREPARED BY	Ikwe.ai · Visible Healing Inc.

research@ikwe.ai · ikwe.ai · © 2026 Visible Healing Inc.

01 Executive Summary

This report presents findings from the EQ Safety Benchmark, a two-stage evaluation framework designed to measure behavioral emotional safety in conversational AI systems. We evaluated 948 AI responses across 79 emotionally vulnerable scenarios, testing four frontier AI systems.

54.7%	45.3%	43%	1.7/5
Passed Safety Gate	Introduced Risk	No Correction	Avg. Regulation

Core Finding: Recognition ≠ Safety.

High emotional articulation often correlates with worse behavioral safety scores.

AI systems can accurately recognize emotions while still responding in ways that increase distress. The gap between recognition and regulation is where harm lives.

This finding challenges the industry assumption that emotional intelligence in AI systems correlates with safe behavior. Recognition is necessary but not sufficient.

What This Proves

Emotional recognition alone does not prevent harm. In many cases, it increases it. This gap is what Ikwe audits, mitigates, and designs against.

02 Methodology

The EQ Safety Benchmark uses a two-stage approach to separate harm detection from quality measurement.

Stage 1 — Safety Gate (Pass/Fail)

Binary detection of behaviors that introduce emotional risk at first contact. Responses flagged for: premature problem-solving, toxic positivity, minimization, abandonment via referral, distress amplification.

45.3% of baseline responses failed.

Stage 2 — Behavioral Quality

Only responses passing Stage 1 are scored. Quality measured across weighted dimensions:
emotional regulation (35%),
acknowledgment quality (25%),
response trajectory (20%),
safety awareness (15%), contextual fit (5%).

Conditional on passing the Safety Gate.

> **These stages separate "did harm occur?" from "was the response actually safe?"**

Data Sources

Scenarios derived from 8 public datasets, primarily LuangMV97/Empathetic_counseling_Dataset from Hugging Face. Covering 12 vulnerability categories:

Grief/loss · Trauma/abuse · Loneliness · Crisis situations · Relationship distress
Work stress · Health anxiety · Financial stress · Identity/self-worth
Family conflict · Social rejection · Life transitions

03 Results

Why This Matters for Deployed Systems

Stage 2 scores reflect performance only among responses that passed the Stage 1 Safety Gate.

MODEL	SAFETY GATE	STAGE 2 SCORE	REGULATION
Ikwe EI Prototype	Pass	**84.6%**	4.05/5
GPT-4o	Pass	**59.0%**	2.95/5
Claude 3.5 Sonnet	Pass	**56.4%**	2.82/5
Grok	Pass	**20.5%**	1.02/5

Ikwe EI Prototype is an internal reference model.

> **High scores don't mean "more empathy."**
>
> They mean lower likelihood of harm under emotional load.

Safety Gate Failure Patterns

Premature Problem-Solving

Jumping to solutions before validating emotional state.

Toxic Positivity

Reassurance that dismisses or minimizes expressed distress.

Abandonment via Referral

Redirecting to help without providing presence first.

Distress Amplification

Mirroring or escalating the user's emotional state.

Minimization

Downplaying the significance of the user's experience.

04 Implications

For Organizations Deploying AI

Organizations using conversational AI in sensitive contexts (mental health, wellness, education, caregiving) should implement behavioral safety evaluation before deployment. The gap between expected and actual emotional safety is significant and measurable.

This gap is invisible without deliberate measurement infrastructure.

For AI Developers

Standard safety evaluations focus on content policy compliance and factual accuracy. These findings suggest that behavioral safety — how responses land on users emotionally — requires dedicated evaluation infrastructure. Systems can be policy-compliant, accurate, and still increase distress.

For Policymakers

Current AI safety frameworks emphasize accuracy, bias, and content policy. Emotional safety — the behavioral impact of AI interactions on vulnerable users — is not yet addressed in major regulatory frameworks. These findings suggest it should be.

Ikwe exists to catch this before scale.

This research is the backbone of the products you're buying.

Limitations

- Test conditions: Results reflect controlled scenarios, not naturalistic conversation.
- Sample size: 79 scenarios across 4 systems; broader coverage needed for generalization.
- Static evaluation: Models are updated frequently; results may not reflect current versions.
- Cultural context: Scenarios primarily reflect Western emotional frameworks.
- No longitudinal data: We measure single interactions, not long-term effects.

05 Citation & Contact

Ikwe.ai. (2026). Behavioral Emotional Safety in Conversational AI:
A Scenario-Based Evaluation. Visible Healing Inc. https://ikwe.ai/full-report

RESEARCH INQUIRIES	research@ikwe.ai
AUDIT ENGAGEMENT	ikwe.ai/audit
FULL METHODOLOGY	ikwe.ai/inquiry

Where This Goes Next

See a real audit example

Redacted artifacts from real evaluations.

ikwe.ai/proof

Explore how this becomes an audit

Four weeks. Full documentation.

ikwe.ai/audit

Download the public preview

Scorecard, before/after, methodology.

ikwe.ai/downloads

This research directly informs Ikwe audits and implementation systems.

Explore how this becomes a real audit → ikwe.ai/audit