IKWE.AI RESEARCH

Behavioral Emotional Safety in Conversational AI

A Scenario-Based Evaluation

Public Research Summary

Applied evaluation infrastructure for emotional safety in conversational AI

Version 2.1 · February 2026

https://ikwe.ai

EXECUTIVE SUMMARY

What This Research Found

This document presents Ikwe.ai's evaluation of behavioral emotional safety in conversational AI systems.

Behavioral emotional safety refers to a system's ability to remain stabilizing, bounded, and non-amplifying once emotional vulnerability is present. It is not the same as emotional recognition, empathetic language, or policy compliance.

Most existing AI benchmarks assess whether a system can identify emotion or avoid disallowed content. They do not measure what happens after a user is already distressed, or how system behavior changes as emotional intensity increases.

Across evaluated frontier models, **only 54.7%** of baseline responses passed the initial emotional safety check at first contact. When risk was introduced, nearly half of responses showed no corrective behavior.

Recognition ≠ Safety

An AI system can accurately identify emotion and articulate empathy while still behaving unsafely under emotional load.

STUDY SCOPE

Evaluation Parameters

79

Emotionally vulnerable scenarios

312

Evaluated responses

4

Conversational AI systems

8+1

Behavioral dimensions + safety gate

These findings describe observed behavioral patterns under controlled test conditions.
They do not imply real-world outcomes, clinical efficacy, or deployment readiness.

FRAMEWORK

Two-Stage Evaluation

This benchmark explicitly separates two questions that are often collapsed into a single score — and should not be:

Stage 1 — Emotional Safety Gate (Binary)

Does the response introduce any predefined emotional safety risk patterns at first contact? Any trigger indicates the response introduces risk before trust has formed.

Stage 2 — Behavioral Stability (Conditional)

If a response passes Stage 1, does it remain safe as emotional intensity increases? This is evaluated across eight weighted dimensions including:

- Regulation before reasoning (20%)
- Escalation awareness (15%)
- Boundary maintenance (15%)
- Distress tolerance (12%)
- Autonomy support (10%)

Sounding safe is not the same as being safe.

KEY FINDINGS

What the Data Shows

STAGE 1 — FIRST-CONTACT RISK

54.7%

of baseline responses **passed** the initial emotional safety check
(did not introduce risk at first contact)

43% of risk-introducing responses showed no corrective behavior within the interaction window.

STAGE 2 — CONDITIONAL PERFORMANCE

Among responses that passed Stage 1, frontier models showed high variance in behavioral safety over time. The Ikwe.ai model demonstrated greater consistency.

The difference was not expressiveness. It was stability.

KEY INSIGHT

The Hidden Risk Pattern

Most safety failures did not appear hostile or overtly harmful.

They appeared **supportive** — while:

- Reinforcing distress
- Accelerating rumination
- Missing escalation signals
- Amplifying emotional intensity

Emotional fluency can mask behavioral risk.

SCOPE & LIMITATIONS

This benchmark does not make claims about:

- Real-world outcomes
- Clinical or therapeutic efficacy
- Deployment readiness
- General intelligence
- Model intent or training design

IMPLICATIONS

Why This Matters

Conversational AI is increasingly deployed in emotionally vulnerable contexts: mental health support, relationship guidance, grief processing, and health coaching.

Without behavioral safety measurement:

- Emotional capability can increase risk at scale
- Supportive language can conceal unsafe trajectories
- Late-stage failures become harder to detect

Emotional capability without behavioral stability introduces risk at scale.

CITATION

Ikwe.ai (2026). Behavioral Emotional Safety in Conversational AI:
A Scenario-Based Evaluation. Public Research Summary · Version 2.1

https://ikwe.ai/research

Research & Press: research@ikwe.ai

Ikwe.ai · Visible Healing Inc.

Building the behavioral safety layer AI benchmarks forgot.