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

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 over the course of an interaction.

Ikwe.ai developed this benchmark after observing repeated safety failures during applied system testing. Systems that appeared emotionally fluent at first contact frequently introduced risk over time — reinforcing distress, accelerating rumination, or failing to maintain appropriate conversational boundaries.

Across evaluated frontier models, more than half of baseline responses to emotionally vulnerable scenarios triggered at least one emotional safety risk pattern at first contact. When risk was introduced, nearly half of responses showed no corrective behavior within the interaction window.

The Ikwe.ai model was evaluated as a reference implementation. It demonstrated lower behavioral variance and greater stability after passing a baseline emotional safety gate, illustrating that emotional capability and behavioral safety are not inherently coupled.

> **Recognition ≠ Safety**
>
> An AI system can accurately identify emotion and articulate empathy while still behaving unsafely under emotional load.

STUDY SCOPE

79	Emotionally vulnerable scenarios
312	Evaluated responses
4	Conversational AI systems
8+1	Behavioral stability dimensions + baseline safety gate

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

CONTEXT

Why This Benchmark Exists

Ikwe.ai did not begin as a benchmark effort. It began with applied systems designed to support humans in emotionally complex situations, including relationship conflict, distress processing, and vulnerability disclosure.

During applied system testing, Ikwe.ai observed a consistent failure pattern:

> Existing AI safety and emotional intelligence benchmarks could not detect — or explain — the behavioral risks emerging during emotionally loaded interactions.

Systems often passed content and tone-based safety checks while simultaneously exhibiting behaviors that increased emotional risk. These behaviors were subtle, conversationally fluent, and difficult to detect without trajectory-based evaluation.

This research formalizes the evaluation infrastructure Ikwe.ai built to measure those risks.

The Emotional Safety Gap

Most AI safety evaluations focus on first-turn behavior under neutral conditions. They rarely test what happens after emotional trust forms, whether a system maintains boundaries as vulnerability deepens, or how a system responds after it introduces emotional risk.

This benchmark measures **observable behavioral patterns under emotional load** — where conversational AI can either stabilize a user or unintentionally increase harm while appearing supportive.

FRAMEWORK

What This Benchmark Measures

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

Stage 1 — Baseline Emotional Safety Gate (Binary)

Does the response introduce any predefined emotional safety risk patterns at first contact?

Any trigger indicates that the response introduces emotional risk before trust has formed.

Stage 2 — Behavioral Stability Over Time (Conditional)

If a response passes the baseline gate, does it remain safe as emotional intensity increases?

Behavioral stability is evaluated across eight weighted dimensions, including:

- Regulation before reasoning
- Escalation awareness
- Boundary maintenance
- Distress tolerance
- Autonomy support

> **Sounding safe is not the same as being safe.**

METHODOLOGY

Response Collection & Evaluation

Scenario Sources. 79 scenarios curated from eight public datasets representing real-world emotional support requests, including CounselChat, MentalChat16K, and the Empathetic Counseling Dataset.

Response Collection.

- Default system configurations
- No custom safety prompts or behavioral priming
- Frontier models tested via API using standard system context
- All responses evaluated post-hoc using the same rubric

Evaluation Framework. Responses were evaluated using:

1. A binary emotional safety gate
2. Conditional behavioral stability scoring across eight weighted dimensions

Regulation Before Reasoning carries the highest weight.

This public summary intentionally omits scoring thresholds, weights, and scenario text to prevent misuse or reverse engineering.

KEY FINDINGS

First-Contact Risk

54.7%

of baseline responses triggered at least one emotional safety risk pattern at first contact

This indicates that more than half of initial responses to emotionally vulnerable scenarios introduced behavioral risk **before any corrective mechanism was engaged**.

Absent Repair Behavior

43%

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

Failures clustered into three dominant categories: reinforcement of rumination loops, boundary and role confusion, and timing and containment gaps.

Behavioral Stability

Among responses that passed the baseline safety gate, 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

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

These findings describe observed behavior under test conditions, not real-world impact.

IMPLICATIONS

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.