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 Shows

This report presents a scenario-based evaluation of **behavioral emotional safety** in conversational AI systems.

Unlike benchmarks that focus on emotional recognition, policy compliance, or content filtering, this framework evaluates how AI systems behave once emotional vulnerability is present — and whether safety is maintained as interactions unfold.

Across evaluated frontier models, emotionally supportive behavior frequently degraded over time, even when initial responses appeared safe.

The **Ikwe.ai model** demonstrated lower behavioral variance and greater stability after passing a baseline emotional safety gate.

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

STUDY SCOPE

79	Scenarios
312	Evaluated responses
4	AI systems
8+1	Scoring dimensions (8 stability + 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. It began with applied systems.

Early emotionally intelligent AI systems — including Lady's Lady and He Said / She Said — were designed to support humans during moments of vulnerability, conflict, and emotional complexity.

During testing, a consistent pattern emerged:

> Existing AI safety and EQ benchmarks could not explain — or detect — the behavioral risks we were encountering.

This research represents the infrastructure built to measure those risks.

The Emotional Safety Gap

Most benchmarks measure capability under neutral conditions. They test for toxicity, bias, refusal behavior, or emotion labeling.

They rarely test what happens after a user is already hurting.

This benchmark measures observable behavior under emotional load — where conversational AI can stabilize a user or unintentionally increase risk while appearing supportive.

FRAMEWORK

What This Benchmark Measures

This evaluation separates two questions often collapsed into a single score:

Stage 1 — Baseline Emotional Safety Gate (Binary)

Does the response meet baseline emotional safety criteria?

A binary check for ten predefined behavioral risk patterns. Any trigger indicates the response introduces emotional risk at first contact.

Stage 2 — Behavioral Stability Over Time (Conditional)

If it starts safe, does it remain safe as vulnerability deepens?

Applied only to responses that pass Stage 1. Measures behavioral stability across eight weighted dimensions, including:

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

> **Starting safe is not the same as staying 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
- Manually tested systems run in fresh sessions
- 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.

FINDINGS

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 patterns associated with increased emotional risk.

43% of baseline responses showed no corrective behavior within the interaction window

When emotional risk was introduced, failures clustered into three categories: escalation and amplification, 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 behavioral consistency.

> **The difference was not expressiveness — it was stability.**

KEY INSIGHT

Most safety failures did not look hostile or overtly harmful.

They looked supportive — while:

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

> **Fluency can mask behavioral risk.**

SCOPE & LIMITS

What This Benchmark Does Not Claim

This benchmark is intentionally scoped. It does not make claims about:

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

These findings describe behavioral patterns 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.