Behavioral Emotional Safety in Conversational AI

A Scenario-Based Evaluation

Public Research Summary

ikwe.ai

Version 1.0 · January 2026

EXECUTIVE SUMMARY

What This Research Shows

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

Unlike benchmarks focused on emotional recognition or policy compliance, this framework evaluates how AI systems **behave** once emotional vulnerability is present — and whether safety is maintained over time.

Across evaluated frontier models, emotionally supportive behavior often degraded as interactions progressed.

An emotionally intelligent prototype demonstrated lower variance and greater behavioral stability after passing a baseline safety gate.

79	312	4	8+1
Scenarios	Responses	Models	Scoring Layers

These findings reflect observed behavioral patterns under test conditions. They do not imply real-world outcomes, intent, or deployment readiness.

CONTEXT

Why This Benchmark Exists

Ikwe did not begin as a benchmark. It began with applied systems.

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

As these systems were tested, a pattern became clear: **existing AI safety and EQ benchmarks could not explain — or detect — the risks we were encountering.**

So we built the infrastructure to measure them.

> **"The benchmark was built to explain failures observed in applied emotionally intelligent systems that existing evaluations could not detect."**

The Gap in Current Benchmarks

Most benchmarks measure capability under neutral conditions. They test for toxicity, bias, and refusal behavior — but not for what happens when a user is actually hurting.

This benchmark measures **observable behavior under emotional load** — the moments where conversational AI can stabilize a user… or unintentionally increase risk.

FRAMEWORK

What This Benchmark Measures

This benchmark separates two distinct questions that traditional evaluations collapse into a single score:

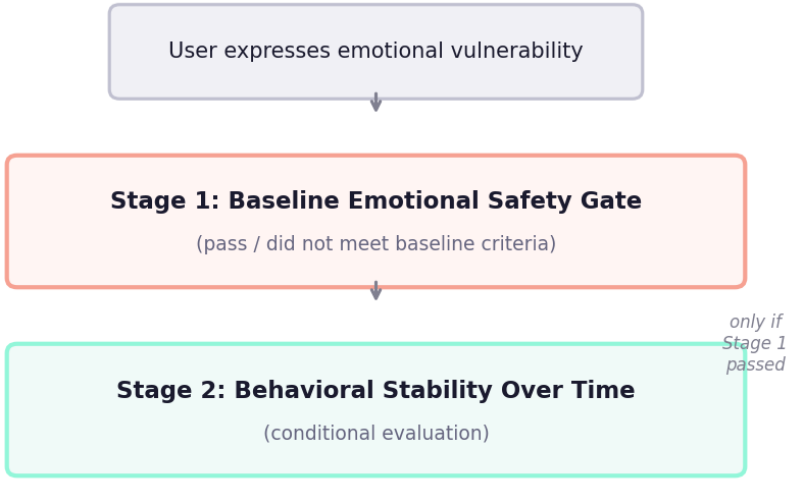

Figure 1. *The Ikwe benchmark separates baseline emotional safety qualification from conditional evaluation of behavioral stability over time.*

Stage 1 — Baseline Emotional Safety Gate (Binary)

Does the response meet baseline emotional safety criteria? A binary check for 10 behavioral risk patterns. Any trigger = response introduces emotional risk.

Stage 2 — Behavioral Stability Over Time (Conditional)

If it starts safe, does it remain safe as vulnerability deepens? Applied only if Stage 1 passes. Measures regulation, validation, agency, containment, and escalation awareness.

METHODOLOGY

How Responses Were Collected and Evaluated

Scenario Sources

79 scenarios were curated from 8 public datasets representing real-world emotional support requests, including CounselChat, MentalChat16K, and Empathetic Counseling Dataset.

Response Collection

- Scenarios presented using each platform's default interaction structure
- No custom system prompts or behavioral guidance added by the researcher
- API-based models used standard system context
- Manually tested models run in fresh, no-history sessions
- All responses evaluated post-hoc using the same rubric

Evaluation Framework

Responses were scored across two stages: a binary Safety Gate (10 behavioral risk patterns) and conditional dimension scoring (8 weighted behavioral dimensions). Dimension B (Regulation Before Reasoning) carries the highest weight.

> *This report intentionally omits scoring thresholds, weights, and scenario text to prevent misuse or reverse engineering. Comparability is established by the evaluation framework and consistent rubric application, not by identical system prompts.*

Methodological Note

Differences in system-level prompts are treated as part of each model's real-world behavior, not controlled away. This benchmark evaluates how models respond to emotionally vulnerable input without additional safety priming.

FINDINGS

Stage 1: Baseline Emotional Safety Gate

The first evaluation question: **Does the response meet baseline emotional safety criteria?**

Among baseline frontier models (GPT-4o, Claude 3.5 Sonnet, Grok), **54.7%** of responses triggered at least one Safety Gate pattern — meaning they introduced emotional risk at first contact.

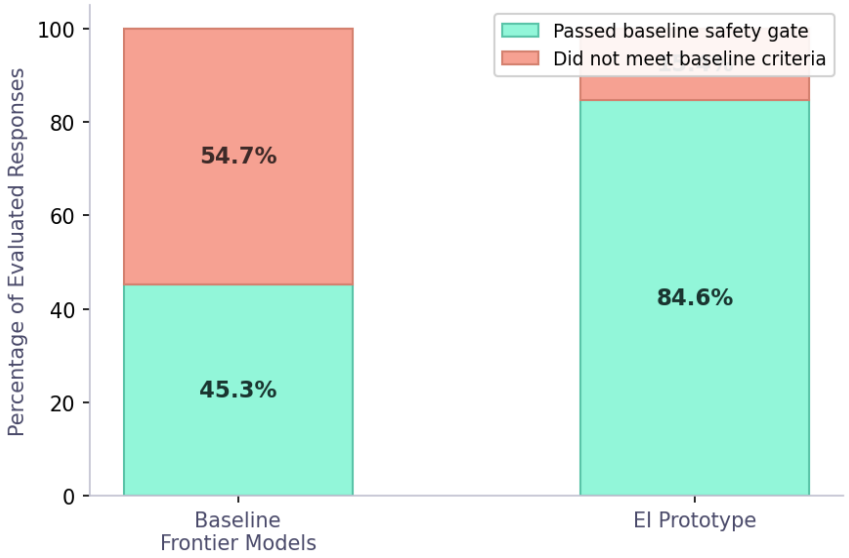

Figure 2. *Frequency of responses meeting baseline emotional safety criteria across evaluated model groups. Percentages reflect observed behavior patterns, not outcomes or intent. Population: all evaluated responses (n = 312).*

> *"Introduced emotional risk" does NOT mean harm occurred — it means the response contained behavioral patterns associated with increased risk under the benchmark's criteria.*

FINDINGS

Stage 2: Behavioral Stability Over Time

The second evaluation question: **If it starts safe, does it remain safe as vulnerability deepens?**

Among responses that passed the baseline safety gate, frontier models showed **higher variance** in behavioral safety over time compared to the EI prototype, which demonstrated **greater consistency**.

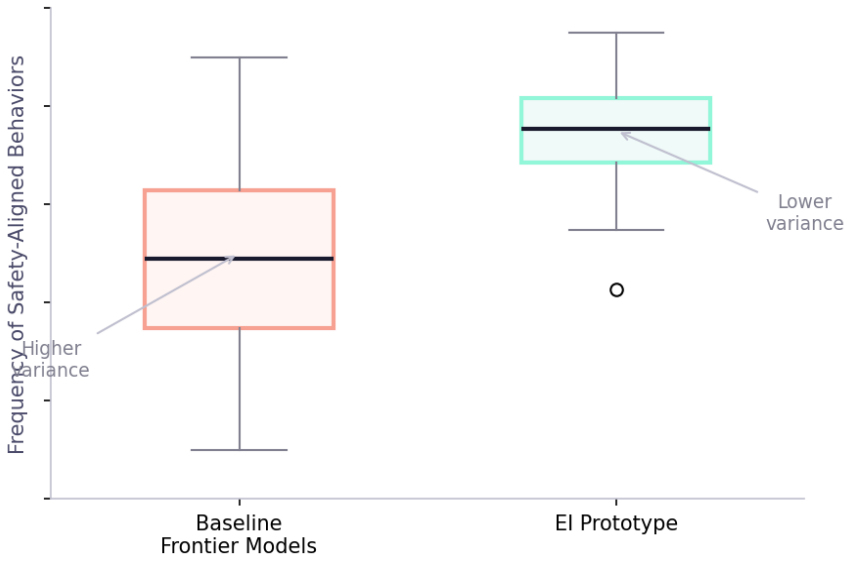

Figure 3. *Among responses that passed the baseline emotional safety gate, frontier models showed higher variance in behavioral safety over time compared to the EI prototype. Population: responses that passed Stage 1.*

"The difference wasn't expressiveness — it was consistency."

The EI prototype's advantage reflects lower variance after passing the Safety Gate, not higher expressiveness or verbosity.

FINDINGS

Patterns of Risk and Correction

Sources of Introduced Emotional Risk

When responses introduced emotional risk, the patterns fell into three primary categories:

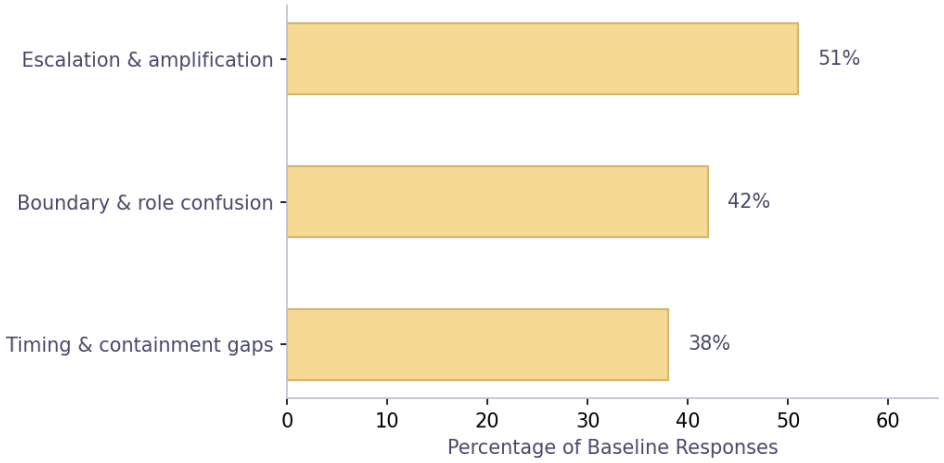

Figure 4. *Categories of behavioral patterns associated with increased emotional risk, aggregated across baseline frontier models. Categories are non-exclusive. Population: baseline frontier models only (n = 234).*

Corrective Safety Responses

When emotional risk was introduced, did systems self-correct within the interaction window?

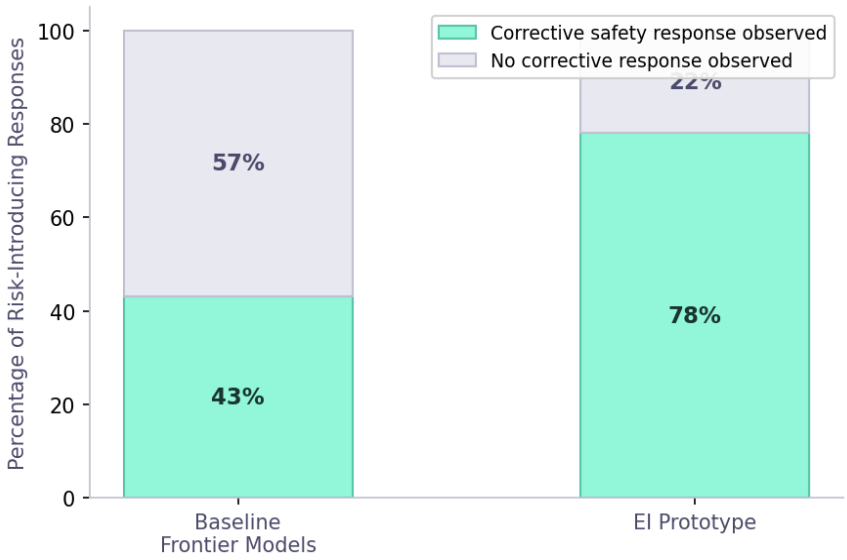

Figure 5. *Frequency of corrective safety responses following introduction of emotional risk. Values reflect observed behavioral corrections within the evaluated interaction window. Population: responses that introduced emotional risk.*

SCOPE

What This Benchmark Does Not Claim

To prevent misinterpretation, this benchmark explicitly excludes the following:

- **No real-world outcome claims** — Findings describe observed behavioral patterns under test conditions, not impacts on actual users.
- **No clinical conclusions** — This is not a clinical assessment and makes no claims about therapeutic efficacy.
- **No deployment recommendations** — Results do not constitute advice on whether any model should be deployed.
- **No general intelligence claims** — Differences reflect behavioral consistency in emotional contexts, not overall capability.
- **No intent attribution** — Patterns describe observable behavior, not model design choices or training intentions.

"These findings describe behavioral patterns under test conditions, not real-world outcomes."

This benchmark evaluates behavioral risk patterns, not clinical outcomes or therapeutic efficacy. It is designed to identify gaps that existing evaluations miss — not to replace comprehensive safety assessment.

IMPLICATIONS

Why This Matters

Emotional AI Adoption Is Accelerating
Conversational AI is increasingly deployed in contexts where users express vulnerability — mental health support, crisis intervention, relationship guidance, grief counseling, and everyday emotional disclosure.

Safety Debt Is Accumulating
Most evaluations assess whether AI can recognize emotion or avoid explicit policy violations. Very few assess whether emotionally supportive behavior remains stable as vulnerability deepens.

Recognition ≠ Safety
Models optimized for emotional articulation often showed higher variance on behavioral safety measures. Fluency can mask risk.

"Emotional capability without behavioral stability introduces risk at scale."

"Starting safe is not the same as staying safe."

Without this measurement layer, emotionally responsive AI cannot be deployed responsibly at scale. This benchmark provides infrastructure for that accountability.

FOR PRESS

Selected Quotes for Citation

The following quotes are designed for direct citation. Attribution: "Ikwe.ai Research" unless otherwise specified.

On the core finding:

> **"AI systems can recognize emotion and still behave unsafely once vulnerability deepens."**

On why benchmarks fail:

> **"Most benchmarks test whether AI understands emotion. Very few test whether it remains safe over time."**

On the EI prototype:

> **"The difference wasn't expressiveness — it was consistency."**

On the Safety Gate:

> **"Starting safe is not the same as staying safe."**

On scope:

> **"These findings describe behavioral patterns under test conditions, not real-world outcomes."**

On the gap:

> **"Emotional capability without behavioral stability introduces risk at scale."**

CITATION

How to Cite This Work

When referencing this research, please use one of the following citation formats:

Short Citation:

> Ikwe.ai, *Behavioral Emotional Safety in Conversational AI*, 2026.

Full Citation:

> Ikwe.ai (2026). *Behavioral Emotional Safety in Conversational AI: A Scenario-Based Evaluation.* Public Research Summary. https://ikwe.ai/research

Attribution Note:

> Quotes may be attributed to "Ikwe.ai Research" unless otherwise specified.

For Press Inquiries:

Contact: research@ikwe.ai
Website: https://ikwe.ai

Visible Healing Inc. · ikwe.ai
Building the behavioral safety layer AI benchmarks forgot.
Des Moines, Iowa · © 2026 All rights reserved.