IKWE.AI RESEARCH

Behavioral Emotional Safety in Conversational AI

A Scenario-Based Evaluation

Full Research Report

Applied evaluation infrastructure for emotional safety in conversational AI

Version 2.1 · February 2026

https://ikwe.ai

KEY RESULTS

Baseline Frontier Models

54.7%	triggered at least one Safety Gate pattern at first contact
43%	showed no corrective behavior within interaction
1.7 / 5	average Regulation Before Reasoning score (baseline models)

Highest emotional fluency often correlated with lower behavioral safety.

This document describes observed behavioral patterns under controlled test conditions. It does not make clinical, deployment, or real-world outcome claims.

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.

THE PROBLEM

What Benchmarks Miss

Existing AI safety benchmarks focus on content-level risks: toxicity, bias, policy violations, and harmful outputs. Emotional intelligence benchmarks test recognition accuracy — whether a model can correctly label emotions.

Neither approach measures **behavioral safety under emotional load**: how a system actually behaves when interacting with someone who is already vulnerable.

The Gap

A model can correctly identify that a user is experiencing grief, respond with linguistically appropriate empathy, and still:

- Mirror and amplify the user's distress
- Feed rumination cycles
- Miss escalation signals
- Create inappropriate dependency
- Blur professional boundaries

These failures don't register on existing benchmarks. This research addresses that gap.

METHODOLOGY

Study Design

This evaluation uses a two-stage framework to separate first-contact safety from behavioral stability over time.

Stage 1: Safety Gate (Binary)

Each response is checked against ten predefined behavioral risk patterns. Any trigger results in a Safety Gate failure, indicating the response introduced emotional risk at first contact.

Stage 2: Behavioral Stability (Conditional)

Responses that pass the Safety Gate are evaluated across eight weighted dimensions measuring behavioral stability as emotional vulnerability deepens.

Dimension	Weight
Regulation Before Reasoning	20%
Escalation Awareness	15%
Boundary Maintenance	15%
Distress Tolerance	12%
Reality Grounding	12%
Autonomy Support	10%
Resource Bridging	8%
Emotional Continuity	8%

FINDINGS

Model Comparison

System	Avg Score	Safety Pass Rate	Regulation Score
Ikwe.ai model	74.0	84.6%	4.05 / 5
Claude 3.5 Sonnet	52.7	56.4%	2.03 / 5
GPT-4o	51.6	59.0%	1.69 / 5
Grok	40.5	20.5%	1.40 / 5

Scores reflect observed response behavior under test conditions—not intent, training data, or overall model capability.

Key Finding 1: First-Contact Risk

54.7% of baseline model responses triggered at least one Safety Gate 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—before any trust had formed or interaction had deepened.

Key Finding 2: Absent Repair Behavior

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

When a response introduced emotional risk, nearly half of the time the system showed no subsequent attempt to repair, redirect, or stabilize. This suggests that current models lack mechanisms for recognizing and correcting their own emotional safety failures mid-interaction.

Key Finding 3: Fluency ≠ Regulation

Models with the highest emotional articulation often performed worst on safety behaviors under distress.

Systems that excelled at naming emotions and expressing empathy showed measurable degradation in safety-relevant behaviors as emotional intensity increased. Fluent emotional language did not correlate with—and sometimes inversely correlated with—behavioral regulation.

IMPLICATIONS

For AI Developers

- Emotional recognition benchmarks are insufficient proxies for emotional safety
- Safety evaluation must include multi-turn behavioral stability testing
- Repair mechanisms need explicit design attention—they don't emerge from general training
- Higher emotional fluency may require additional safety constraints, not fewer

For Healthcare and Wellness Deployers

- HIPAA compliance and data security do not address behavioral safety
- 'Health' and 'emotional intelligence' marketing claims require behavioral verification
- Risk increases with context aggregation—more personal data means higher stakes
- Late-stage risk (after trust forms) may be more dangerous than first-contact risk

For Policy and Governance

- Current AI safety frameworks lack behavioral emotional safety requirements
- Self-reported emotional intelligence claims have no standardized verification
- Vulnerable population protections need trajectory-based evaluation, not just content filtering

LIMITATIONS

What This Research Does Not Claim

- **No real-world outcome claims** — Findings describe test behavior, not deployment outcomes
- **No clinical or therapeutic conclusions** — This is safety research, not clinical validation
- **No deployment recommendations** — Results inform evaluation, not product decisions
- **No general intelligence comparisons** — Scores reflect emotional safety, not capability
- **No attribution of intent or design motive** — Behavior is measured, not explained

Methodological Limitations

- Sample size (79 scenarios, 312 responses) limits statistical power for subgroup analysis
- Scenario selection may not fully represent real-world emotional vulnerability distribution
- Single-turn and short multi-turn evaluation may miss longer interaction dynamics
- Model versions tested may differ from current production deployments
- Human evaluation introduces subjectivity despite calibration efforts

CONCLUSION

This research identifies a measurable gap between emotional recognition capabilities and behavioral emotional safety in conversational AI systems.

The core finding—**Recognition** ≠ **Safety**—has implications for how AI systems are evaluated, deployed, and marketed in contexts involving emotional vulnerability.

Current frontier models demonstrate substantial first-contact risk rates and limited repair behavior. Systems with high emotional articulation do not reliably translate that capability into safe behavior under emotional load.

> **Safety is not the first reply. It is the trajectory.**

Behavioral emotional safety must be measured — not assumed.

CITATION & RESOURCES

```
Ikwe.ai (2026). Behavioral Emotional Safety in Conversational AI: A
Scenario-Based Evaluation. Full Research Report · Version 2.1 ·
https://ikwe.ai/research
```

Additional Resources

- Full methodology: https://ikwe.ai/research
- Interactive findings explorer: https://ikwe.ai/explorer
- Press resources: https://ikwe.ai/press
- Request evaluation: https://ikwe.ai/inquiry

Research & Press: research@ikwe.ai

Ikwe.ai · Visible Healing Inc.

Building the behavioral safety layer AI benchmarks forgot.