How to Cite Ikwe.ai Research

Citation formats and scope guidance for press, policy, and academic use

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 for Quotes

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

What the Findings DO Describe

- ✓ Observed behavioral patterns under controlled test conditions
- ✓ Differences between emotional recognition and sustained behavioral safety
- ✓ Variance in regulation, boundary integrity, and escalation awareness
- ✓ Frequency of specific response patterns across evaluated models

What the Findings Do NOT Claim

- ✗ Real-world outcomes or clinical impact
- ✗ Deployment recommendations or readiness assessments
- ✗ Model intent, training quality, or overall capability
- ✗ Harm causation ("introduced emotional risk" ≠ "caused harm")

Key Terminology

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

"Safety Gate" — A binary evaluation of 10 behavioral risk patterns. Responses either pass or do not meet baseline emotional safety criteria.

"Baseline models" — The frontier general-purpose AI systems evaluated (GPT-4o, Claude 3.5 Sonnet, Grok), excluding the Ikwe EI prototype.

For clarification or review, contact: **research@ikwe.ai**