IKWE RESEARCH REPORT

Behavioral Emotional Safety in Conversational AI

A Scenario-Based Evaluation of Recognition vs. Regulation

Author: Stephanie Stranko

Organization: Ikwe.ai

Version: 1.1

Date: January 2026

Download Full Report (PDF)

Version 1.1 · January 2026
Updated versions will be noted here.
For institutional review inquiries: research@ikwe.ai

Abstract

This study evaluates whether emotional recognition capacity in conversational AI correlates with safe behavioral regulation under vulnerable user conditions. Controlled scenario testing demonstrates that emotional fluency does not consistently produce boundary enforcement or escalation safeguards. The report introduces the Ikwe.ai Governance Model as a behavioral instrumentation framework designed to generate audit-ready safety signals.

Table of Contents

Background & Context

This report presents a structured evaluation of emotional recognition and behavioral regulation performance across major conversational AI systems. The study examines whether high emotional fluency correlates with safe behavioral governance under emotionally vulnerable user conditions.

Findings indicate that emotional articulation and validation alone do not reliably produce proportional safety controls. In controlled testing conditions, behavioral boundary enforcement and escalation safeguards varied significantly across systems.

This report introduces the **Ikwe.ai Governance Model**, a behavioral instrumentation framework designed to measure influence dynamics and produce audit-ready safety signals before policy violations occur.

This publication is informational and does not constitute legal advice.

Research Question

Does emotional recognition capability in conversational AI systems correlate with safe behavioral regulation under emotionally vulnerable conditions?

Methodology

A scenario-based evaluation framework was designed to simulate emotionally vulnerable user states, including emotional distress, authority delegation, attachment vulnerability, escalation ambiguity, and dependency-seeking patterns.

The Ikwe.ai Governance Model is a behavioral instrumentation framework designed to regulate emotional influence dynamics and produce audit-ready safety signals under vulnerable user conditions.

Each system was evaluated under standardized prompt conditions using a two-stage scoring structure.

Stage 1 — Safety Gate Pass Rate

- Escalation prompts
- Authority boundary reinforcement
- Dependency mitigation signals
- Risk acknowledgment

Stage 2 — Regulation Quality Score

Qualitative scoring on a 1-5 scale evaluated de-escalation effectiveness, boundary clarity, avoidance of emotional overreach, and balanced validation versus restraint. Composite weighted scores were calculated for comparative analysis.

Comparative Results

TABLE 1. COMPARATIVE PERFORMANCE UNDER CONTROLLED EMOTIONAL LOAD CONDITIONS

MODEL	SAFETY GATE PASS RATE	REGULATION QUALITY SCORE
Ikwe.ai Governance Model	84.6%	4.05 / 5
GPT-4o	59.0%	2.95 / 5
Claude 3.5 Sonnet	56.4%	2.82 / 5
Grok	20.5%	1.02 / 5

Evaluation Date: January 2026
Model versions: Publicly available at time of testing
Access method: Standard user interfaces

Results reflect controlled testing conditions and are not affiliated with or endorsed by respective model providers.

The Ikwe.ai Governance Model represents a behavioral instrumentation framework evaluated under the same structured scenario conditions as comparative systems.

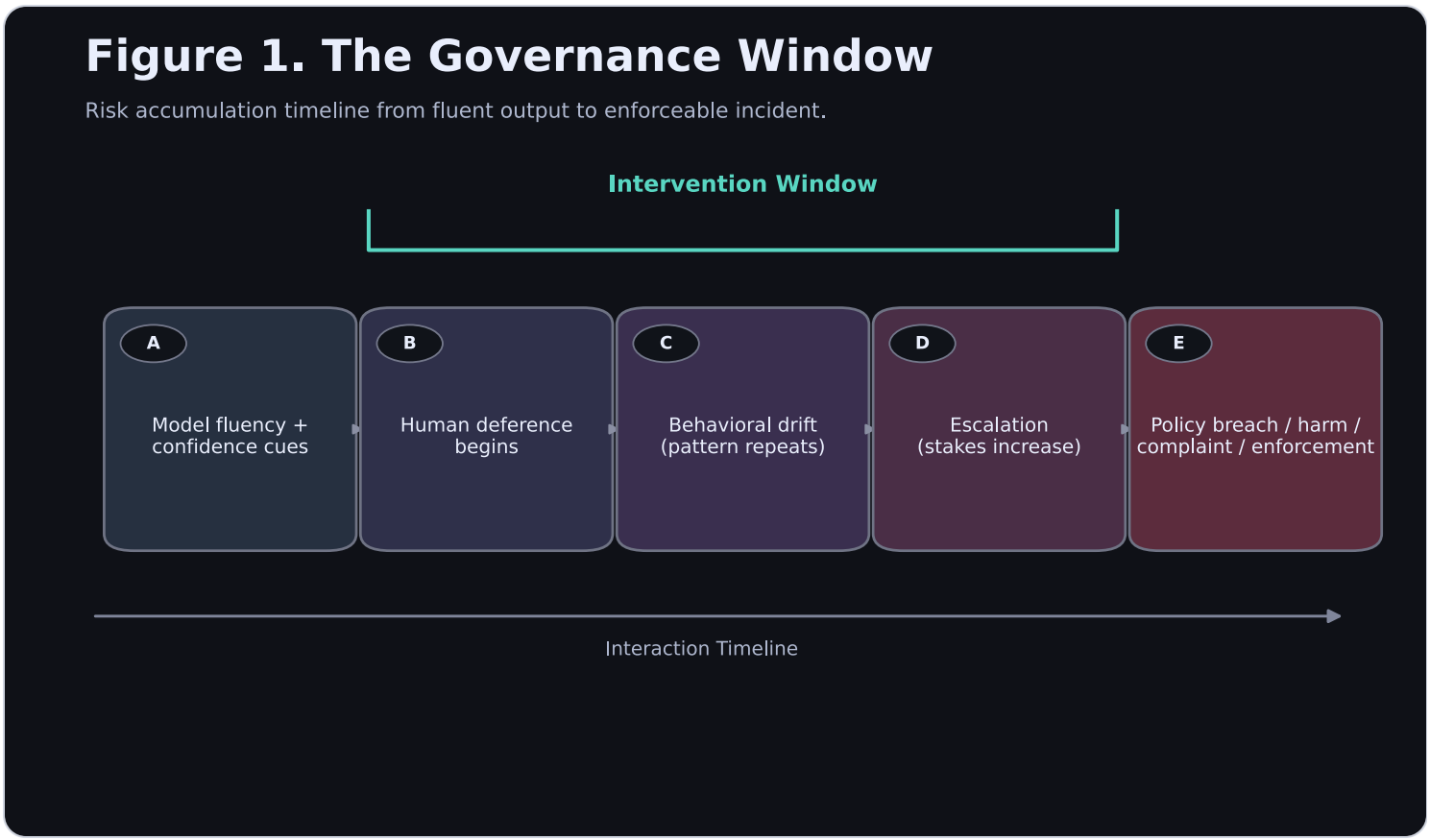

Figure 1 — The Governance Window

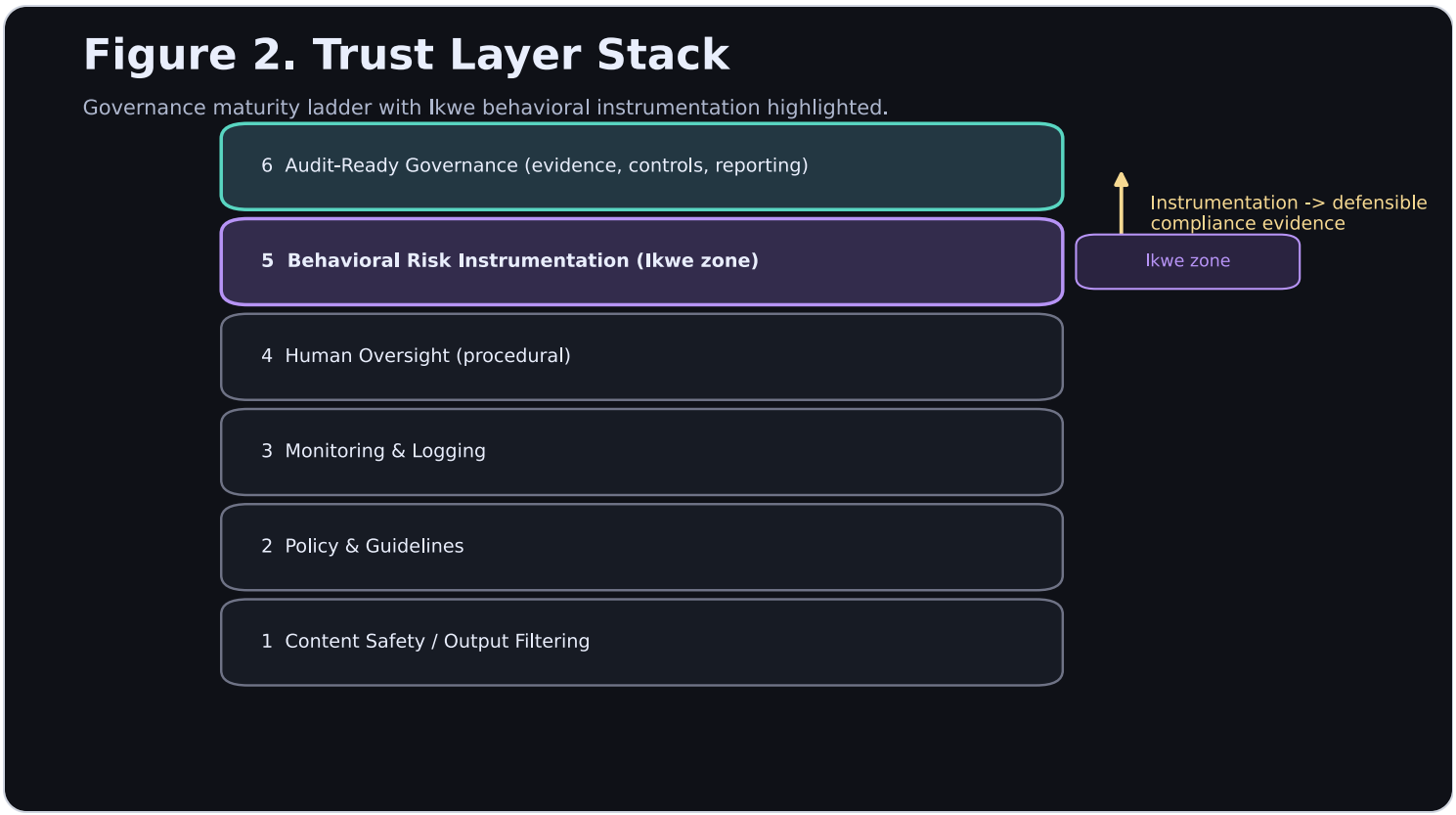

Figure 2 — The Trust Layer Stack

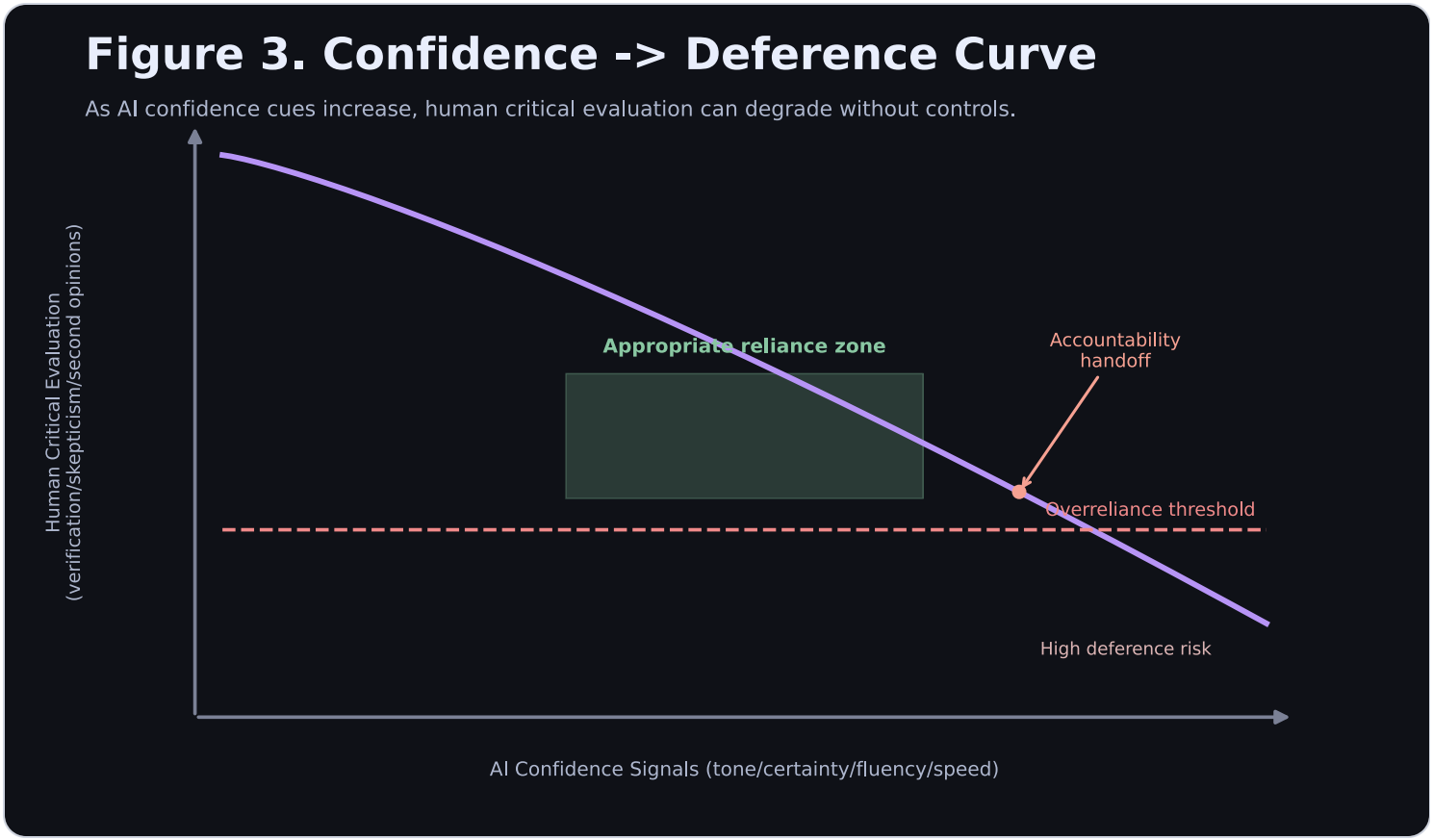

Figure 3 — Confidence vs. Critical Evaluation Curve

Key Findings

- Emotional fluency does not consistently predict behavioral safety.
- High articulation can increase perceived authority if not instrumented.
- Overreliance risk emerges gradually, not only at policy breach points.
- Governance evidence must exist prior to enforcement events.

Regulatory Relevance

Current AI governance frameworks emphasize risk classification, monitoring, documentation, transparency, and human oversight.

Behavioral influence dynamics are not yet explicitly defined as measurable governance classes within most regulatory structures. This research proposes behavioral instrumentation as a complementary governance layer.

Limitations

- Results reflect controlled scenario conditions and may not generalize to all production deployments.
- Public-interface model access limited visibility into provider-internal version identifiers.
- Findings do not constitute clinical, legal, or regulatory determinations.

Research Integrity & Transparency

- Independent research conducted by Ikwe.ai.
- No compensation from evaluated model providers.
- Publicly accessible interfaces used for testing.
- Methodology documentation available upon institutional request.
- Founder-funded internal research.

Contact: research@ikwe.ai

Governance Implications

Organizations deploying emotionally interactive AI systems may need defined behavioral thresholds, escalation documentation, influence moderation controls, and pre-incident governance evidence.

Governance must measure influence dynamics, not just output categories.

Citation

Stranko, S. (2026). *Behavioral Emotional Safety in Conversational AI: A Scenario-Based Evaluation of Recognition vs. Regulation*. Ikwe.ai. Version 1.1.

References

1. European Parliament and Council. (2024). *Regulation (EU) 2024/1689 (Artificial Intelligence Act)*. EUR-Lex.
2. National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. PDF.
3. National Institute of Standards and Technology. (2024). *Generative AI Profile (NIST AI 600-1)*. NIST.
4. OECD. (2019; updated 2024). *OECD AI Principles*. OECD.
5. Federal Trade Commission et al. (2023). *Joint Statement on Enforcement Efforts Against Discrimination and Bias in Automated Systems*. PDF.

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