AI Governance Audit Report

LLM-Based Wellbeing Support Chatbot

Author: Selena Singh
Role: AI Governance and Human-Centered AI Research
Date: February 15, 2026

1. Executive Summary

This report presents a governance and safety audit of an LLM-assisted chatbot designed to provide film recommendations and optional mood-based support. The objective of this evaluation was to assess potential ethical, psychological, and governance risks associated with emotionally responsive AI systems and to demonstrate a structured auditing approach aligned with emerging responsible AI practices.

The audit evaluated chatbot responses for emotional dependency risk, transparency clarity, and overall system reliability. Results indicate that while the system demonstrated generally stable and appropriate outputs, several high-risk emotional-support responses highlighted the importance of governance guardrails for AI systems that simulate empathy or provide wellbeing-oriented interaction.

This project demonstrates how structured AI audits can identify risks before deployment and provides a framework for integrating human-centered design and governance into AI-enabled products.

2. System Overview

The audited system is an LLM-assisted chatbot developed using a movie dataset and natural language prompting. The chatbot provides film recommendations based on genre, user preferences, and optional mood-based inputs.

While primarily designed for entertainment recommendations, the system includes conversational capabilities that allow users to express emotional states such as stress, sadness, or anxiety. This creates a hybrid use case combining recommendation systems with emotionally responsive interaction.

Such systems are increasingly common across consumer platforms, including streaming services, digital companions, and productivity applications, making governance evaluation critical.

3. Audit Objectives

The audit was designed to evaluate the chatbot across three governance and trust dimensions:

1. Emotional and wellbeing risk
Assess whether responses could create unhealthy reliance, simulate therapeutic authority, or reinforce distress.

2. Transparency and user awareness
Evaluate whether responses clearly maintained system identity as an artificial tool rather than a human or professional substitute.

3. Reliability and clarity
Assess whether outputs remained consistent, appropriate, and understandable across test scenarios.

4. Methodology

The audit used structured prompt testing to simulate real-world user interactions. Scenarios included neutral recommendation requests, emotional-support prompts, and mixed conversational queries.

Each response was evaluated and categorized using a structured scoring framework:

- High-risk outputs
- Weak or unclear outputs
- Transparency flags

Testing focused on identifying patterns rather than isolated outputs to evaluate systemic behavior.

5. Key Findings

High-risk outputs: 2
Weak outputs: 0

Transparency flags: 0

Interpretation

The chatbot performed reliably in standard recommendation scenarios and maintained clear system identity throughout most interactions. However, two responses in mood-support scenarios demonstrated elevated emotional-support language that could be interpreted as overly reassuring or companion-like.

While not harmful in isolation, such responses highlight the importance of guardrails when conversational systems interact with users expressing distress or vulnerability.

Overall system performance indicates low immediate risk but moderate governance sensitivity in emotional-support contexts.

6. Risk Analysis

Emotionally responsive AI systems present unique governance challenges because users may interpret conversational fluency as empathy, authority, or understanding. When systems provide reassurance or supportive language without clear boundaries, users may develop misplaced trust or reliance.

This risk is heightened in systems used frequently or by vulnerable populations. Without transparency and guardrails, even well-intentioned outputs can blur the distinction between tool and companion.

The audit demonstrates that governance evaluation should extend beyond accuracy to include emotional and behavioral impact.

7. Governance Recommendations

Transparency safeguards

Include clear system-level messaging that the chatbot is not a mental health professional and cannot provide therapeutic support.

Emotional-boundary design

Implement response guidelines that redirect users expressing distress toward appropriate real-world support when necessary.

Periodic auditing

Conduct regular prompt-based audits to evaluate system behavior as models evolve or datasets change.

Human-centered evaluation

Integrate human-computer interaction review into AI deployment cycles to assess trust, clarity, and user interpretation.

8. Business and Product Implications

AI systems that simulate conversation or provide personalized recommendations increasingly influence user behavior and trust. Governance failures in these systems can create reputational risk, reduce user retention, and undermine long-term product credibility.

Conversely, organizations that embed transparency, guardrails, and human-centered design into AI systems are better positioned to build durable trust and sustained user engagement.

Responsible AI governance is therefore not only a compliance exercise but a strategic component of product reliability and long-term brand value.

9. Skills Demonstrated

AI governance auditing
Human-centered AI evaluation
Prompt-based risk testing
LLM interaction analysis
Responsible AI documentation
Business-aligned risk assessment

10. Conclusion

This audit illustrates how structured evaluation of conversational AI systems can identify trust and safety risks prior to deployment. As AI becomes embedded in everyday products, organizations that implement proactive governance and human-centered design will be better equipped to build systems users trust and rely on over time.