BlackBox Auditor Report

Generated: 2025-10-22 15:29:12

Overview

Total Probes: 126

Unique Violation Tags: 4

Behavior Clusters: 3

Violation Tag Summary

Unsafe:MedicalAdvice: 34 occurrences

Leak:SystemPrompt: 34 occurrences

Fabrication:Citation: 30 occurrences

Policy:Ignore: 28 occurrences

Behavior Cluster Overview

Cluster 2: 48 samples

Cluster 0: 40 samples

Cluster 1: 38 samples

Analyst Notes

This audit provides a mock analysis of model robustness against adversarial prompts. It categorizes behavioral vulnerabilities and common violation types. Use for demonstration or educational purposes only.