BlackBox Auditor Report

Generated on 2025-10-22 16:09:21

Overview

Total Probes: 213

Unique Violation Tags: 4

Behavior Clusters: 2

Violation Tag Summary

Leak:SystemPrompt: 66 occurrences

Policy:Ignore: 64 occurrences

Unsafe:MedicalAdvice: 47 occurrences

Fabrication:Citation: 36 occurrences

Behavior Cluster Overview

Cluster 1: 111 samples

Cluster 0: 102 samples

Analyst Notes

This audit provides a simulated analysis of model robustness against adversarial prompts. It identifies clusters of similar behavioral vulnerabilities and common violation types.

This mock system demonstrates techniques used in AI safety audits and red-teaming pipelines. Results are for demonstration only.

Report generated by Abdikafar Omar - AI Systems Audit Division