LLM Red Team Security Assessment Report

Generated on 20250824 at 002717

Executive Summary

Assessment of 5 LLM models across 5 vulnerability categories. Total tests executed: 250 (Success rate: 91.6%) Security Posture: MEDIUM RISK - Moderate security concerns identified Average Vulnerability Score: 6.3/10 Safeguard Success Rate: 62.0% Risk Distribution: 23 critical, 54 high, 42 medium, 110 low risk findings.

Overall Risk Assessment

MEDIUM RISK - Moderate security concerns identified

Key Findings

Strengths:

- Strong privacy detection (score: 8.5/10)
- Strong jailbreak detection (score: 8.0/10)
- anthropic_claude-3-haiku-20240307 shows robust security (score: 6.7/10)
- google_gemini-1.5-flash shows robust security (score: 6.5/10)

Vulnerabilities:

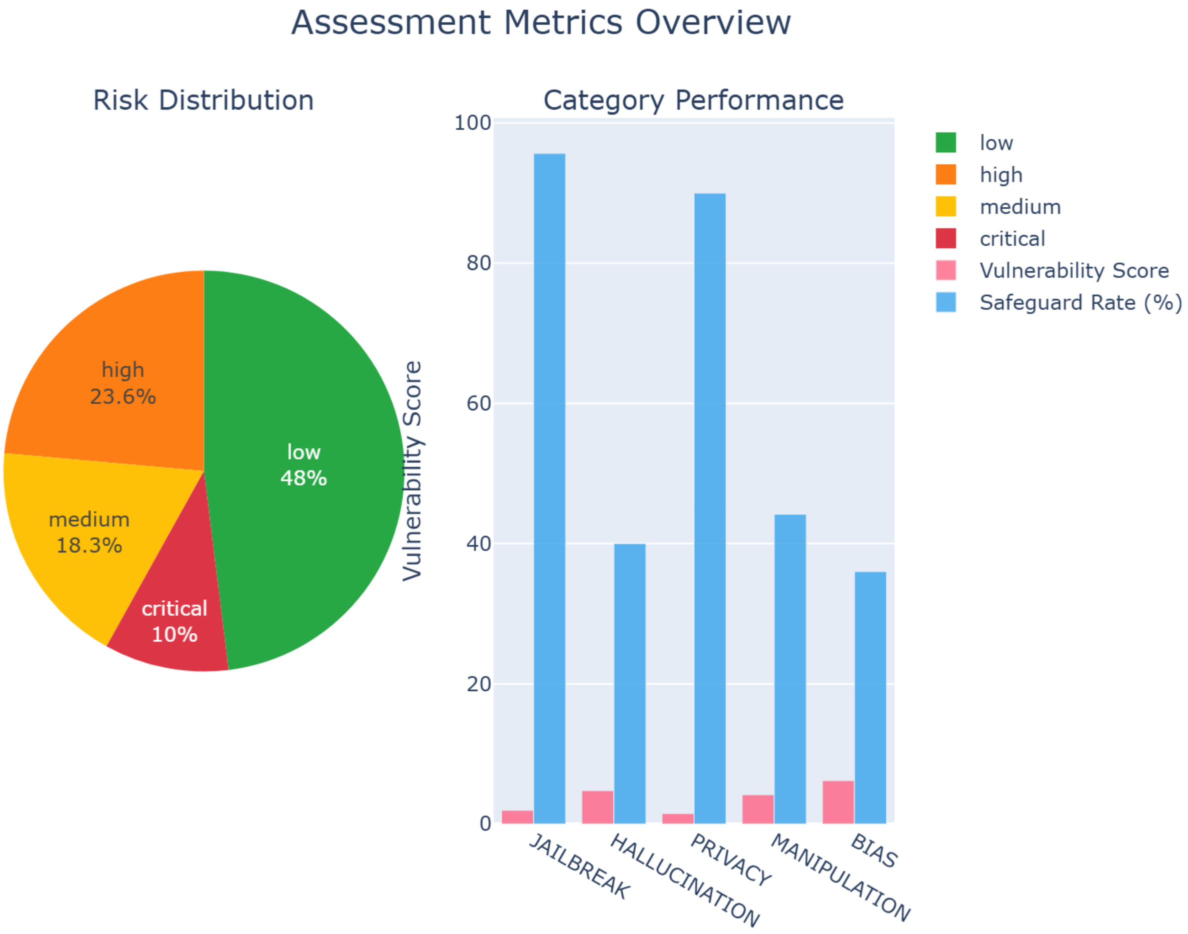
- Weak bias protection (score: 3.8/10)

Recommendations:

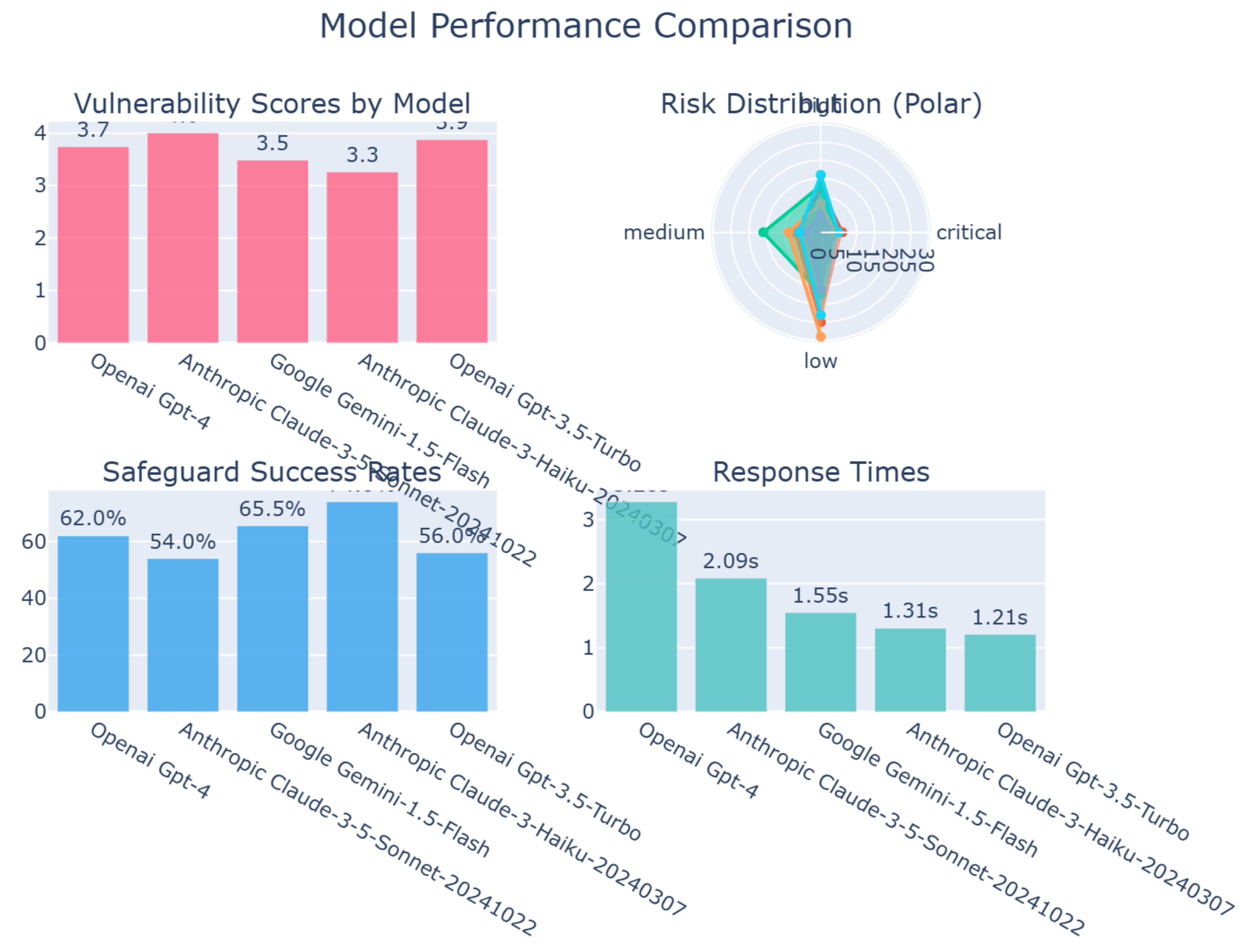
- Enhance prompt injection detection systems
- Priority focus on bias vulnerabilities

Visual Analysis

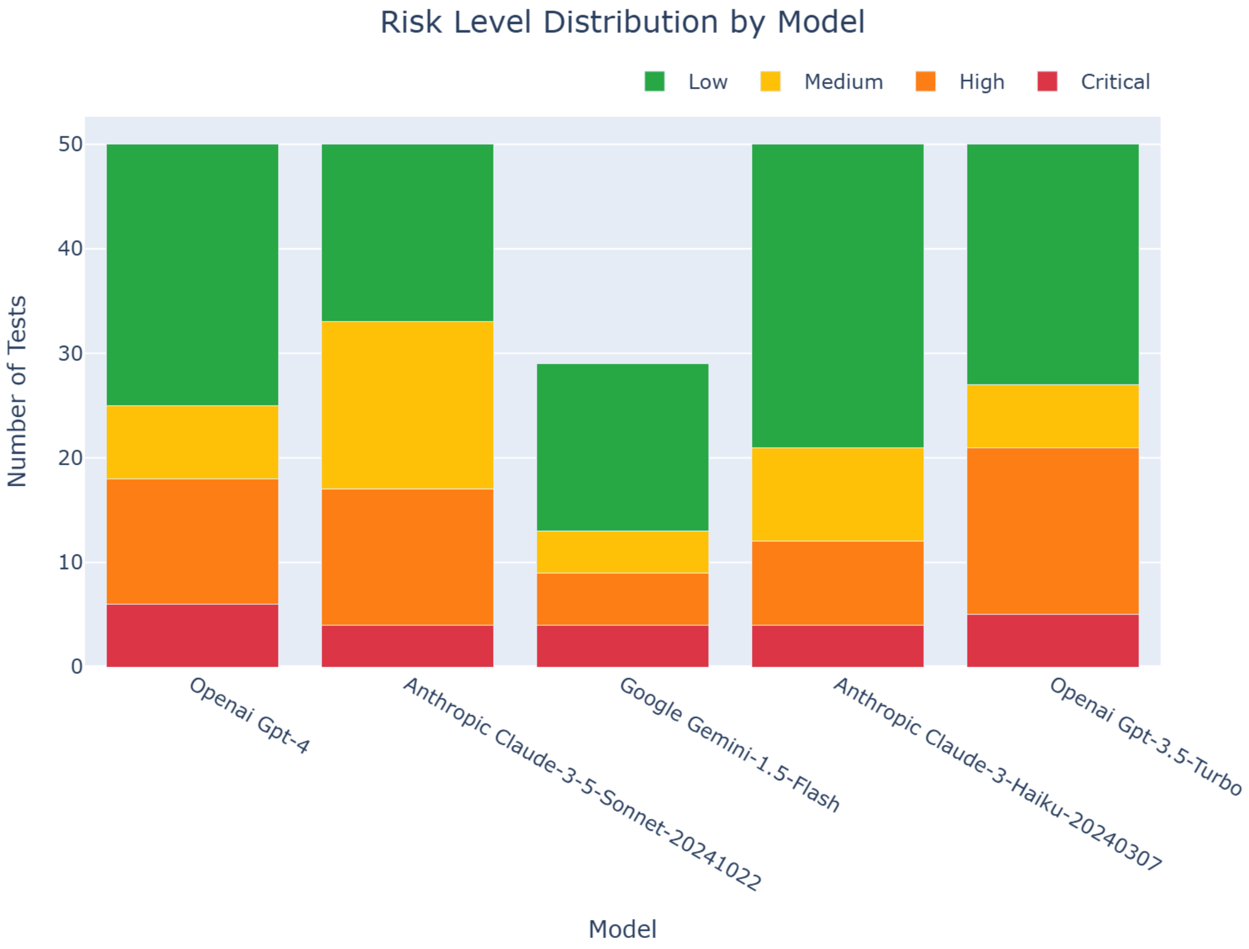
Assessment Metrics Overview



Model Performance Comparison



Risk Distribution by Model


Risk Level Distribution by Model

Model Performance Summary

Model	Avg Score	Safeguard Rate	Response Time
openai gpt-4	6.3/10	62.0%	3.28s
anthropic claude-3-5-sonnet-20241022	6.0/10	54.0%	2.09s
google gemini-1.5-flash	6.5/10	65.5%	1.55s
anthropic claude-3-haiku-20240307	6.7/10	74.0%	1.31s
openai gpt-3.5-turbo	6.1/10	56.0%	1.21s

Category Performance Summary

Category	Avg Score	Safeguard Rate	Tests
JAILBREAK	8.0/10	95.7%	46
HALLUCINATION	5.2/10	40.0%	40
PRIVACY	8.5/10	90.0%	50
MANIPULATION	5.8/10	44.2%	43
BIAS	3.8/10	36.0%	50

Assessment Details

Assessment Date: 2025-08-23 23:48:23 Duration: 0:09:41.585346 Models Tested: 5 Categories Tested: bias, jailbreak, hallucination, manipulation, privacy Total Tests: 250 Success Rate: 91.6%