AI Safety Audit Report

Sample Demonstration Report

Report Type:	Comprehensive AI Safety Audit
AI System:	Customer Support Chatbot (Demo)
Audit Date:	March 07, 2026
Models Tested:	25+ LLM Models
Attack Strategies:	17 Reframing Techniques
Compliance Framework:	EU AI Act Article 15

Executive Summary

This sample report demonstrates UMMRO's comprehensive AI safety auditing capabilities. In a production audit, this section would contain a high-level overview of findings, risk assessment, and strategic recommendations for securing your AI systems.

Key Findings

Category	Risk Level	Findings	Recommendation
Prompt Injection	MEDIUM	3 of 25 models vulnerable	Implement input sanitization
Jailbreak Attempts	LOW	1 of 25 models vulnerable	Apply constitutional AI
Data Leakage	CRITICAL	5 of 25 models at risk	Immediate remediation required
Bias & Fairness	MEDIUM	Detected in 8 models	Implement bias detection
EU AI Act Compliance	HIGH	12 requirements met	Address 3 gaps identified

Testing Methodology

UMMRO employs a research-backed testing framework that evaluates AI systems across 17 distinct reframing strategies and 25+ leading LLM models. Our methodology includes:

- Automated prompt injection testing across multiple attack vectors
- Jailbreak attempt detection using adversarial prompting techniques
- Soft refusal detection and terminology sanitization analysis
- Multi-model orchestration for comprehensive coverage
- EU AI Act Article 15 compliance verification
- Boardroom-ready PDF reports with actionable recommendations

Model Coverage (Sample)

Model Provider	Model	Tested	Risk Score
OpenAI	GPT-4	✓	2.3/10
Anthropic	Claude 3.5 Sonnet	✓	1.8/10
Google	Gemini Pro	✓	3.1/10
Meta	Llama 3.1	✓	4.2/10

Mistral	Mistral Large	✓	3.7/10

EU AI Act Compliance Mapping

UMMRO automatically maps audit findings to EU AI Act Article 15 requirements, helping organizations demonstrate compliance with emerging AI regulations.

Requirement	Status	Evidence
Transparency obligations	MET	Model cards documented
Risk management system	MET	Audit findings logged
Data governance	PARTIAL	2 gaps identified
Technical documentation	MET	Full system documentation
Human oversight measures	GAP	Requires implementation

Next Steps & Recommendations

- **1. Immediate Actions:** Address CRITICAL findings within 48 hours
- **2. Short-term (1-4 weeks):** Implement recommended safeguards for MEDIUM-risk items
- **3. Long-term (1-3 months):** Establish ongoing AI safety monitoring program
- **4. Compliance:** Close identified EU AI Act gaps before regulatory deadlines
- **5. Training:** Conduct staff training on AI safety best practices