Adversarial Testing Report

Executive Summary

This report presents the results of adversarial testing conducted on July 13, 2025. The testing evaluated the system's resilience against various attack vectors including confidentiality breaches, integrity violations, inference attacks, and toxicity incidents.

Test Configuration

Parameter	Value
AUT Model	qwen/qwen3-8b
Testing Model	qwen/qwen3-14b
Max Rounds	4
Timeout	3600 seconds
Account Data Path	C:\Users\Sanyam\Desktop\San\instance\uploads\data.json

Test Results Overview

Category	Tests Conducted	Status
Confidentiality	0	Completed
Integrity	0	Completed
Inference	0	Completed
Toxicity	0	Completed
Total Tests	0	Completed

Detailed Test Results

No detailed test results available.

Recommendations

Based on the adversarial testing results, the following recommendations are provided: 1. **Data Protection:** Implement stricter data access controls and validation mechanisms. 2. **Response Filtering:** Enhance output filtering to prevent sensitive information leakage. 3. **Consistency Checks:** Implement consistency validation across multiple interactions. 4. **Toxicity Prevention:** Strengthen content moderation and response safety measures. 5. **Continuous Monitoring:** Establish regular adversarial testing schedules.

Files Generated

This report is accompanied by the following detailed files: • Prompts File: test_prompts_20250713_021756.json • Responses File: aut_responses_20250713_021756.json • Analysis Report: security_report_20250713_021756.json Report generated on: 2025-07-13 02:18:23