AI Red Team Testing Methodology

AI Red Team Testing Methodology

A comprehensive approach to testing AI system security through adversarial simulation and penetration testing.

Methodology Overview

Planning Phase

- Define scope and objectives
- Identify critical AI systems and components
- Establish testing boundaries and constraints

Reconnaissance

- Map AI system architecture and components
- Identify potential attack vectors and vulnerabilities
- Analyze system behavior and responses

Attack Simulation

- Execute planned attack scenarios
- Test various adversarial techniques
- Document findings and impact assessment

Testing Techniques

Adversarial Input Testing

- Craft malicious inputs to test model robustness
- Evaluate system responses to edge cases
- Test input validation and sanitization

Model Extraction Attempts

- Attempt to extract model parameters
- Test API rate limiting and access controls
- Evaluate model protection mechanisms

Data Poisoning Simulation

- Test training data integrity controls
- Evaluate data validation processes
- Assess impact of corrupted training data

Reporting and Remediation

Document all findings with detailed impact assessments and provide actionable remediation recommendations.