Focused POC Test Plan for LLM Security Tools

Limited to 2 Scans Per Tool

Overview

This streamlined plan is designed for a rapid proof of concept (POC) evaluation of Protect AI Recon, Virtue AI, and Hiddenlayer with the constraint of only 2 scans per tool. The plan prioritizes high-impact test scenarios to maximize evaluation effectiveness within these limitations.

Prerequisites (Already Completed)

- Cross-functional team is assembled
- Demo accounts for all three tools are available
- Basic understanding of each tool's interface is established

Test Environment Setup (1-2 Days)

- Create a controlled test environment with:
- One open-source LLM (e.g., Llama 2)
- One commercial API (e.g., GPT-4)
- Non-sensitive but representative healthcare data structures
- Document baseline performance metrics
- Configure each security tool according to vendor documentation

Scan 1: Healthcare Security Fundamentals (All Tools)

Objective

Evaluate each tool's effectiveness in detecting fundamental healthcare security vulnerabilities.

Test Cases

1. **PHI Protection Test Suite**
2. Craft 5-10 prompts attempting to extract patient data
3. Test variations of PHI extraction attempts (names, DOB, diagnoses, etc.)
4. Document detection rates and alert quality
5. **Basic Prompt Injection**
6. Test 3-5 standard prompt injection techniques
7. Include healthcare-specific injection scenarios
8. Evaluate detection accuracy and response time
9. **HIPAA Compliance Verification**
10. Test against core HIPAA Security Rule requirements
11. Evaluate quality of compliance reporting
12. Document remediation guidance quality

Metrics to Collect

- Detection rate (% of vulnerabilities identified)
- False positive rate
- Alert quality and actionability
- Remediation guidance clarity
- Performance impact on LLM response time

Scan 2: Advanced Healthcare Security (All Tools)

Objective

Evaluate each tool's capabilities for advanced security scenarios and tool-specific strengths.

Common Test Cases

1. **Advanced Jailbreak Attempts**
2. Test 3-5 sophisticated jailbreak techniques
3. Include healthcare-specific scenarios (e.g., unauthorized medical advice)
4. Document detection effectiveness

5. **Data Leakage Scenarios**

6. Test for model training data extraction
7. Attempt to extract sensitive healthcare information
8. Evaluate protection mechanisms

Tool-Specific Test Cases

Protect AI Recon

- Test AWS/cloud integration capabilities
- Evaluate guardrail implementation recommendations

Virtue AI

- Test multimodal capabilities (if applicable to your use case)
- Evaluate regulatory compliance features

Hiddenlayer

- Test one-click vulnerability assessment
- Evaluate OWASP LLM alignment

Metrics to Collect

- Detection rate for advanced attacks
- Quality of tool-specific features
- Integration effectiveness
- Reporting comprehensiveness
- Overall security posture improvement

Evaluation Framework (1 Day)

Quantitative Assessment

For each tool, score the following on a 1-5 scale: - Detection effectiveness - False positive/negative rate - Ease of use - Quality of reporting - Remediation guidance - Healthcare-specific capabilities - Integration potential

Qualitative Assessment

Document observations on: - User experience - Learning curve - Quality of alerts - Actionability of findings - Support responsiveness - Fit with CVS Health workflows

Timeline

- Day 1: Environment setup and tool configuration
- Day 2: Scan 1 - Healthcare Security Fundamentals (all tools)
- Day 3: Analysis of Scan 1 results
- Day 4: Scan 2 - Advanced Healthcare Security (all tools)
- Day 5: Final analysis and recommendation development

Critical Success Factors

1. **PHI Protection**: >95% detection of PHI extraction attempts
2. **Compliance Reporting**: Comprehensive HIPAA-aligned reporting
3. **False Positives**: <10% false positive rate
4. **Integration**: Minimal disruption to existing workflows
5. **Actionability**: Clear, implementable remediation guidance

Documentation Requirements

For each scan, document: 1. Test scenarios executed 2. Tool configuration details 3. Detection results (with screenshots) 4. False positives/negatives 5. Performance impact 6. Unique observations

Final Deliverables

1. Comparative scorecard of all three tools
2. Specific strengths and limitations observed
3. Recommendation for CVS Health implementation
4. Implementation considerations for selected tool(s)

Maximizing Value from Limited Scans

Preparation Tips

- Thoroughly review tool documentation before testing
- Prepare all test cases in advance
- Create templates for consistent documentation
- Ensure all team members understand evaluation criteria

Execution Tips

- Run identical test cases across all tools for fair comparison
- Document results in real-time
- Capture screenshots of significant findings
- Note any unexpected behaviors or limitations

Analysis Tips

- Normalize results for fair comparison
- Consider both security effectiveness and operational impact
- Evaluate against CVS Health's specific requirements
- Document both quantitative metrics and qualitative observations

This focused approach will provide meaningful evaluation results despite the constraint of only two scans per tool, enabling CVS Health to make an informed decision based on actual performance in your environment.