ClawGuard Shield

AI Agent Security Compliance Report

10

Risk Score (0-10)

Severity: CRITICAL

Organization: Anthropic MCP

Report Generated: 2026-04-02 20:38 UTC

Findings: 16 threat(s) detected

Scan Time: 15663ms

Scanner: ClawGuard Shield v0.5.0 (50 patterns, 7 categories)

EU AI Act Compliance Reference Included (Enforcement: 02 August 2026)

1. Executive Summary

The scan detected 16 security threat(s) with an overall risk score of 10/10 (CRITICAL). This indicates critical security concerns that require attention. The scan completed in 15663ms using deterministic pattern matching across 42 attack vectors.

Scan Statistics

Total Findings	**16**
Risk Score	**10 / 10**
Overall Severity	**CRITICAL**
Scan Duration	**15663ms**
Patterns Checked	**42**
Attack Categories	**5**
Detection Method	**Deterministic Pattern Matching**
False Positive Rate	**0%**

Severity Breakdown

2. Detailed Findings

Output Injection (6 finding(s))

Attempts to inject malicious content (XSS, SQL injection) into AI agent outputs. OWASP LLM02: Insecure Output Handling.

HIGH 85% **Template Injection (LLM05)**

Line: 0

Match: ${{ steps.find-packages.outputs.packages }} steps: - uses: actions/checkout@v...

Recommendation: Server-side template injection (SSTI) payload. Can lead to remote code execution. OWASP LLM05.

HIGH 85% **Command Injection in Output (LLM05)**

Line: 0

Match: ` to continue.', '', 'If this PR is adding a new server, please close it and ...

Recommendation: Command injection via backticks or $() subshell. If output is rendered in shell context, this executes. OWASP LLM05.

HIGH 85% **Template Injection (LLM05)**

Line: 0

Match: ${{ steps.find-packages.outputs.packages }} steps: - uses: actions/checkout@v...

Recommendation: Server-side template injection (SSTI) payload. Can lead to remote code execution. OWASP LLM05.

HIGH 85% **Command Injection in Output (LLM05)**

Line: 47

Match: `Invalid resourceType: ${args?.resourceType}. Must be ${RESOURCE_TYPE_TEXT} o...

Recommendation: Command injection via backticks or $() subshell. If output is rendered in shell context, this executes. OWASP LLM05.

HIGH 85% **Command Injection in Output (LLM05)**

Line: 59

Match: `Invalid resourceId: ${args?.resourceId}. Must be a finite positive integer.`

Recommendation: Command injection via backticks or $() subshell. If output is rendered in shell context, this executes. OWASP LLM05.

HIGH 85% **Command Injection in Output (LLM05)**

Line: 79

Match: `This prompt includes the ${resourceType} resource with id: ${resourceId}. Pl...

Recommendation: Command injection via backticks or $() subshell. If output is rendered in shell context, this executes. OWASP LLM05.

Shell Injection (2 finding(s))

CRITICAL 90% **Bash Command Substitution $(...)**

Line: 23

Match: $(find

Recommendation: CRITICAL: Bash command substitution $() detected with a shell command.

CRITICAL 90% **Bash Command Substitution $(...)**

Line: 22

Match: $(find

Recommendation: CRITICAL: Bash command substitution $() detected with a shell command.

Data Exfiltration (5 finding(s))

Techniques to steal sensitive data through the AI agent by embedding hidden requests, markdown image injections, or encoded payloads that transmit data to external servers.

MEDIUM 65% **Email Address (LLM06)**

Line: 92

Match: actions@github.com

Recommendation: Email address detected. OWASP LLM06: Sensitive Information Disclosure. DSGVO -- personal identifier.

MEDIUM 70% **Email Address (LLM06)**

Line: 8

Match: jadamson@anthropic.com

Recommendation: Email address detected. OWASP LLM06: Sensitive Information Disclosure. DSGVO -- personal identifier.

MEDIUM 70% **Email Address (LLM06)**

Line: 8

Match: davidsp@anthropic.com

Recommendation: Email address detected. OWASP LLM06: Sensitive Information Disclosure. DSGVO -- personal identifier.

MEDIUM 65% **Email Address (LLM06)**

Line: 8

Match: mariusz@korzekwa.dev

Recommendation: Email address detected. OWASP LLM06: Sensitive Information Disclosure. DSGVO -- personal identifier.

MEDIUM 65% **Email Address (LLM06)**

Line: 637

Match: john@example.com

Recommendation: Email address detected. OWASP LLM06: Sensitive Information Disclosure. DSGVO -- personal identifier.

Code Obfuscation (3 finding(s))

Obfuscated code patterns (Base64, hex encoding, string concatenation, dynamic imports) used to hide malicious payloads from static analysis.

CRITICAL 90% **Python subprocess/os.system**

Line: 42

Match: subprocess.run(

Recommendation: CRITICAL: Direct OS command execution via Python subprocess/os module.

CRITICAL 90% **Python subprocess/os.system**

Line: 101

Match: subprocess.run(

Recommendation: CRITICAL: Direct OS command execution via Python subprocess/os module.

CRITICAL 90% **Python subprocess/os.system**

Line: 107

Match: subprocess.run(

Recommendation: CRITICAL: Direct OS command execution via Python subprocess/os module.

3. Remediation Priorities

The following remediation steps are ordered by severity. Address CRITICAL and HIGH findings immediately before deploying the AI agent to production.

1. CRITICAL **Bash Command Substitution $(...)**
   CRITICAL: Bash command substitution $() detected with a shell command.

2. CRITICAL **Python subprocess/os.system**
   CRITICAL: Direct OS command execution via Python subprocess/os module.

3. HIGH **Template Injection (LLM05)**
   Server-side template injection (SSTI) payload. Can lead to remote code execution. OWASP LLM05.

4. HIGH **Command Injection in Output (LLM05)**
   Command injection via backticks or $() subshell. If output is rendered in shell context, this executes. OWASP LLM05.

5. MEDIUM **Email Address (LLM06)**
   Email address detected. OWASP LLM06: Sensitive Information Disclosure. DSGVO -- personal identifier.

4. EU AI Act Compliance Reference

The EU Artificial Intelligence Act (Regulation 2024/1689) establishes a comprehensive framework for AI systems in the European Union. Full enforcement begins 02 August 2026. The following articles are directly relevant to AI agent security scanning.

Article 9 - Risk Management System

Requirement:

High-risk AI systems require a risk management system throughout the system's lifecycle, including identification and analysis of known and foreseeable risks.

How this scan addresses it:

Prompt injection scanning directly addresses the requirement to identify and mitigate foreseeable security risks in AI systems.

Article 15 - Accuracy, Robustness and Cybersecurity

Requirement:

High-risk AI systems shall be resilient against attempts by unauthorized third parties to alter their use, outputs or performance by exploiting system vulnerabilities.

How this scan addresses it:

Documented prompt injection testing demonstrates compliance with cybersecurity robustness requirements.

Article 17 - Quality Management System

Requirement:

Providers of high-risk AI systems shall put a quality management system in place that ensures compliance, including procedures for data management, risk management, and post-market monitoring.

How this scan addresses it:

Regular security scanning with documented reports forms part of the required quality management system.

Article 61 - Post-Market Monitoring

Requirement:

Providers shall establish and document a post-market monitoring system proportionate to the nature of the AI system.

How this scan addresses it:

Continuous security scanning and compliance reporting satisfies post-market monitoring obligations for AI security.

Important Note

This report provides a security assessment of AI agent inputs. It does not constitute legal advice. EU AI Act compliance requires a comprehensive risk management approach. Consult qualified legal counsel for full regulatory compliance.

5. Methodology

Detection Engine

ClawGuard uses deterministic regex-based pattern matching - not LLM-based detection. This eliminates the fundamental vulnerability of using an LLM to detect attacks against LLMs.

Pattern Coverage

50 attack patterns across 7 categories: Prompt Injection, System Prompt Extraction, Data Exfiltration, Social Engineering, and Context Manipulation.

Performance

All scans complete in under 10ms with zero external API calls. The scanner operates fully offline.

False Positive Rate

Tested against real-world benign content: 0% false positive rate. Patterns are tuned for high precision.

Multilingual Support

Patterns include English and German variants for key attack types.