ClawGuard Shield

AI Agent Security Compliance Report

10

Risk Score (0-10)

Severity: CRITICAL

Organization: Paperclip AI

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

Findings: 28 threat(s) detected

Scan Time: 10062ms

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 28 security threat(s) with an overall risk score of 10/10 (CRITICAL). This indicates critical security concerns that require attention. The scan completed in 10062ms using deterministic pattern matching across 42 attack vectors.

Scan Statistics

Total Findings	**28**
Risk Score	**10 / 10**
Overall Severity	**CRITICAL**
Scan Duration	**10062ms**
Patterns Checked	**42**
Attack Categories	**5**
Detection Method	**Deterministic Pattern Matching**
False Positive Rate	**0%**

Severity Breakdown

2. Detailed Findings

Dangerous Command (11 finding(s))

MEDIUM 65% **Package / Dependency Install**

Line: 30

Match: npm install

Recommendation: Software installation command detected. Verify the package source for supply-chain safety.

MEDIUM 65% **Package / Dependency Install**

Line: 88

Match: npm install

Recommendation: Software installation command detected. Verify the package source for supply-chain safety.

MEDIUM 65% **Package / Dependency Install**

Line: 112

Match: npm install

Recommendation: Software installation command detected. Verify the package source for supply-chain safety.

MEDIUM 65% **Package / Dependency Install**

Line: 150

Match: npm install

Recommendation: Software installation command detected. Verify the package source for supply-chain safety.

MEDIUM 65% **Package / Dependency Install**

Line: 38

Match: npm install

Recommendation: Software installation command detected. Verify the package source for supply-chain safety.

MEDIUM 65% **Package / Dependency Install**

Line: 59

Match: npm install

Recommendation: Software installation command detected. Verify the package source for supply-chain safety.

MEDIUM 65% **Package / Dependency Install**

Line: 54

Match: npm install

Recommendation: Software installation command detected. Verify the package source for supply-chain safety.

MEDIUM 65% **Package / Dependency Install**

Line: 93

Match: npm install

Recommendation: Software installation command detected. Verify the package source for supply-chain safety.

MEDIUM 65% **Package / Dependency Install**

Line: 143

Match: npm install

Recommendation: Software installation command detected. Verify the package source for supply-chain safety.

MEDIUM 65% **Package / Dependency Install**

Line: 181

Match: npm install

Recommendation: Software installation command detected. Verify the package source for supply-chain safety.

MEDIUM 65% **Package / Dependency Install**

Line: 222

Match: npm install

Recommendation: Software installation command detected. Verify the package source for supply-chain safety.

Output Injection (7 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: 9

Match: {{ github.event.pull_request.number }}

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

HIGH 95% **Template Injection (LLM05)**

Line: 26

Match: {{ github.event.pull_request.base.sha }}

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

HIGH 95% **Template Injection (LLM05)**

Line: 26

Match: {{ github.event.pull_request.head.sha }}

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

HIGH 95% **Template Injection (LLM05)**

Line: 85

Match: {{ github.event.pull_request.base.sha }}

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

HIGH 95% **Template Injection (LLM05)**

Line: 85

Match: {{ github.event.pull_request.head.sha }}

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

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

Line: 0

Match: {{ github.token }} run: | if git diff --quiet -- pnpm-lock.yaml; then echo "L...

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

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

Line: 91

Match: $(gh pr list --head chore/refresh-lockfile --json url --jq '.[0].url')

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

Shell Injection (3 finding(s))

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

Line: 48

Match: $(awk

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

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

Line: 56

Match: $(grep

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

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

Line: 64

Match: $(find

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

Data Exfiltration (7 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.

LOW 50% **IP Address in Sensitive Context (LLM06)**

Line: 166

Match: 127.0.0.1

Recommendation: IP address detected. OWASP LLM06. DSGVO -- IP addresses are personal data under EU law.

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

Line: 66

Match: lockfile-bot@users.noreply.github.com

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

HIGH 85% **Password in Cleartext**

Line: 79

Match: PASSWORD=$SMOKE_ADMIN_PASSWORD

Recommendation: Cleartext password detected. Never store or transmit passwords in plain text.

LOW 50% **IP Address in Sensitive Context (LLM06)**

Line: 9

Match: 0.0.0.0

Recommendation: IP address detected. OWASP LLM06. DSGVO -- IP addresses are personal data under EU law.

HIGH 80% **Password in Cleartext**

Line: 6

Match: PASSWORD: paperclip

Recommendation: Cleartext password detected. Never store or transmit passwords in plain text.

HIGH 85% **Database Connection String**

Line: 25

Match: postgres://paperclip:paperclip@db:5432/paperclip

Recommendation: Database connection string with potential credentials detected.

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

Line: 53

Match: embedded-postgres@18.1.0-beta.16.patch

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

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. HIGH **Template Injection (LLM05)**
   Server-side template injection (SSTI) payload. Can lead to remote code execution. OWASP LLM05.
3. HIGH **Command Injection in Output (LLM05)**
   Command injection via backticks or $() subshell. If output is rendered in shell context, this executes. OWASP LLM05.
4. HIGH **Password in Cleartext**
   Cleartext password detected. Never store or transmit passwords in plain text.
5. HIGH **Database Connection String**
   Database connection string with potential credentials detected.
6. MEDIUM **Package / Dependency Install**
   Software installation command detected. Verify the package source for supply-chain safety.
7. MEDIUM **Email Address (LLM06)**
   Email address detected. OWASP LLM06: Sensitive Information Disclosure. DSGVO -- personal identifier.
8. LOW **IP Address in Sensitive Context (LLM06)**
   IP address detected. OWASP LLM06. DSGVO -- IP addresses are personal data under EU law.

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.