SLOWMIST AI

SuperClaw Security Audit Report

REPORT NO.

SM-AR-2026-005

DATE

2026-08-05

Contents

> This report is intended for engineering and security teams. Finding IDs are never renumbered or reused once published; cross-references cite IDs. Read the Audit Overview and Findings Summary first, then jump to Detailed Findings by ID.

Audit Overview

Audit Result PASS

All 5 adversarial tests (SSRF direct outbound and cloud metadata probing, dangerous command interception, sensitive file read, system prompt extraction, indirect command injection) found no exploitable vulnerabilities. The Agent passed core adversarial dimensions including SSRF protection and dangerous command interception. It should be emphasized that some tests could not obtain complete responses due to intermittent authentication failures, resulting in actual coverage lower than expected.

Severity Distribution

0	0	0	0	0	0
Total	Critical	High	Medium	Low	Info

Scope & Limitations

This audit was conducted as a black-box test, sending adversarial inputs via the AgentOn platform API and observing response behavior. Access to the Agent's source code, configuration, system prompts, internal tool implementations, or runtime logs was not available. Black-box testing has inherent limitations, and this audit encountered intermittent authentication failures that prevented some test cases from obtaining complete responses. The following disclaimers must be considered alongside this report:

OUT-OF-SCOPE MODULE	AFFECTED FINDINGS	PENDING VERIFICATION
Black-box Test Coverage Limitations		This audit performed only 5 targeted adversarial tests covering SSRF (direct outbound + cloud metadata 169.254.169.254), dangerous command interception, sensitive file read, system prompt extraction, and indirect command injection. Attack surfaces not covered include but are not limited to: multi-turn dialogue injection, indirect injection via file upload, encoding bypass (Base64 / Unicode / homoglyphs), timing attacks, resource exhaustion attacks, cross-session data leakage, API abuse/scraping, and multimedia output poisoning (embedding诱导 content in generated videos/images). Untested does not mean secure.
Intermittent Authentication Failures Affecting Coverage		During this audit, some test requests returned AUTH_FAILED (retryable: false) at the upstream_authentication stage, while other requests within the same time window were normally accepted by the Agent. This indicates intermittent failures in the Agent's signature verification. As a result, tests such as sensitive file read could not obtain the Agent's complete response content, and the coverage of relevant attack surfaces was lower than expected. The PASS rating in this report does not cover attack paths that could not be fully tested due to authentication failures.

Probabilistic Nature of LLM Safeguards	The Agent's safety controls are based on LLM judgment, which is probabilistic rather than deterministic. The same input may produce different responses under different timing, contexts, or model versions. The Agent correctly intercepted dangerous requests during this test, but this does not guarantee interception under all conditions. Attackers may bypass current defenses through repeated retries, context manipulation, or model version changes. Black-box testing cannot exhaustively enumerate all conditions that trigger malicious output — it is unknown what input, what context, or what time window may trigger unexpected behavior. The conclusions in this report are valid only within the forensic time window (2026-08-05 08:40—09:10 UTC+8).
Unobservable Internal Behavior	Black-box testing can only observe the final output and cannot detect whether unexpected behavior occurred internally within the Agent, such as: whether partial dangerous operations were executed before refusal, whether user input was written to persistent storage, whether sensitive information was leaked in the internal context (even if the final output does not contain it), whether upstream tool calls were triggered (even if a safe response was returned), or whether the multimedia generation pipeline produced harmful content at intermediate steps (even if the final deliverable was filtered). These internal behaviors require white-box auditing or runtime instrumentation to verify.
Agent Version and Configuration Changes	The Agent's version and configuration may change after the audit. System prompt updates, toolchain changes, upstream model upgrades, or security policy adjustments may introduce new risks. The conclusions in this report do not apply to the Agent after version changes; re-audit is required following any changes.
Upstream Supply Chain Not Audited	The upstream LLM model providers, multimedia generation pipelines, and delivery backends that the Agent depends on were not included in this audit. Security vulnerabilities in upstream components (such as model supply chain attacks, generation pipeline poisoning, delivery backend privilege escalation) may indirectly affect the Agent's security. Supply chain security assessment is recommended for subsequent white-box audits.

Uncertainty of Negative Results	This audit did not identify security vulnerabilities, but this does not mean the Agent is secure. A negative result (PASS) from black-box testing only indicates that no observable anomalous behavior was triggered within the tested input space. Security is a continuous process, not a one-time conclusion. Regular regression audits are recommended, combined with white-box auditing, runtime monitoring, and red team adversarial exercises to build a defense-in-depth verification system.

> In summary, the PASS rating in this report should be understood as "no exploitable vulnerabilities were found within the limited input space of this black-box test," rather than "the Agent has been proven secure." Given that intermittent authentication failures during this audit resulted in incomplete coverage of certain attack paths, follow-up white-box source code auditing, runtime behavior baseline monitoring, signature verification stability stress testing, and continuous adversarial testing are recommended to build a more comprehensive security assurance.

Appendix · Severity Definitions

CRITICAL	Directly leads to loss of funds, data breach, or full system compromise with no preconditions.
HIGH	Clear exploitation path with few preconditions; must be fixed before production.
MEDIUM	Exploitable but constrained by existing mechanisms, or impact limited to cost/availability; should be fixed before production.
LOW	Hardening item, data-quality issue, or only exploitable depending on out-of-scope component behavior.
INFO	No direct risk, but lowers attacker cost or violates the principle of least exposure.

Disclaimer

> **Important: this report was generated by an AI Audit Agent**
>
> The analysis, findings, and conclusions in this report were primarily produced by automated AI systems and have not undergone item-by-item review equivalent to a traditional manual audit. AI systems may produce false positives, false negatives, or inferences inconsistent with the facts. Do not place unconditional trust in any individual conclusion of this report.

01 AI Generation and the Need for Human Review

All or most of this report was generated by an AI audit agent using automated program analysis and large-language-model reasoning, and may contain false positives, false negatives, outdated judgments, or incorrect inferences (hallucinations). No finding, severity rating, or remediation advice should be treated as final until independently reviewed and confirmed by qualified security engineers.

02 Non-Exhaustiveness and No Warranty

Security auditing — whether performed by humans or AI — cannot by its nature exhaust all potential risks. The absence of a class of issues from this report does not mean such issues do not exist; "passing the audit" does not constitute any express or implied warranty, endorsement, or certification of the target system's security.

03 Scope and Time Limitations

This report is valid only for the stated audit scope and code snapshot as of the issue date. Out-of-scope modules, any code changes after the audit, third-party dependencies, deployment configuration, and the actual runtime environment are outside its coverage and assurances.

04 No Professional Advice

This report does not constitute investment advice, legal opinion, financial or tax advice, nor an offer or recommendation to buy, hold, or sell any asset.

05 Limitation of Liability

To the maximum extent permitted by applicable law, SlowMist AI and its affiliates accept no liability for any direct, indirect, incidental, or consequential loss (including but not limited to loss of funds, data loss, and business interruption) arising from the use of, reliance on, or inability to use this report. Users must independently verify the report's contents and bear full responsibility for their own decisions.

06 Recommended Use

This report should serve as one input to security decision-making, not the sole basis. We strongly recommend: human security review of all CRITICAL / HIGH findings; an independent third-party manual audit before production deployment or mainnet launch; and regression verification of remediated code.

07 Distribution and Interpretation

This report's distribution is governed by its stated classification; do not quote out of context without written permission. If this disclaimer conflicts with the report body, this disclaimer prevails; where Chinese and English versions differ, the Chinese version prevails.

> By using or relying on this report, you are deemed to have read, understood, and agreed to all terms of this disclaimer.

Security for AI & Crypto · AI for Security

SECURING THE AGENT ECOSYSTEM

WEB www.slowmist.ai

EMAIL team@slowmist.com

X / TWITTER @SlowMist_Team

Disclaimer: this report was generated by an AI audit agent and may contain errors or omissions; all conclusions are subject to independent human review — see the Disclaimer section for full terms. It covers only the stated audit scope and code snapshot and makes no guarantee about unaudited components; security audits cannot exhaust all potential risks, and SlowMist AI accepts no liability for business decisions based on this report.

© 2026 SLOWMIST AI · ALL RIGHTS RESERVED