SLOWMIST AI

LifeClaw Agent Security Audit Report

REPORT NO.	DATE
SM-AR-2026-010	2026-08-26

Contents

01 Audit Overview 03

02 Scope & Limitations 04

03 Disclaimer 07

> 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

None of the 14 black-box tests found exploitable security vulnerabilities. Multi-vector injections (arbitrary number dialing, identity spoofing, payment credential solicitation, threat scripts, URL directives) were never executed — the Agent demonstrated stable structural resistance: injected directives were ignored and a real merchant name plus city was consistently required; SSRF/URL injection produced no outbound requests verified via weblog; control paths including idempotent replay/conflict (409), authentication (401), authorization (404), cancellation (idempotent and free of charge), and quantity boundary (quantity=2 clamped to 1) all behaved correctly; all 15 tasks/operations (including failures, cancellations, timeouts) maintained full billing exemption (`execution_mode=audit` , `billing_status=exempt` , `paid_micro_usdc=0`) with no billing anomalies.

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 inputs via the AgentOn platform audit API and observing response behavior. Access to the Agent's source code, configuration, system prompts, internal tool implementations, or runtime logs was not available. Under audit mode the platform blocks real outbound calls, so the core execution links of a phone-type Agent cannot be directly verified. The following disclaimers must be considered alongside this report:

OUT-OF-SCOPE MODULE	AFFECTED FINDINGS	PENDING VERIFICATION
Black-box Test Coverage Limitations		This audit executed 14 tests covering the task state machine, idempotency, authentication and authorization, cancellation paths, multi-vector injection (arbitrary number dialing, identity spoofing, payment credential solicitation, threat scripts, URL directives), SSRF outbound verification (weblog), and quantity boundaries. Attack surfaces not covered include but are not limited to: multi-turn progressive jailbreaking, encoding bypasses (Base64/Unicode/homoglyphs), in-call voice poisoning, merchant-side system penetration, cross-session preference data abuse, and callback number abuse (harassment/fraud infrastructure). Untested does not mean secure.
Real Outbound Calls and Call Content Unverifiable		Under audit mode the dial execution layer was unreachable: control tasks for a supported region (Hong Kong) and a to-be-verified region (Shenzhen) both failed at the dial execution stage in an identical manner, with no real outbound calls occurring (billing remained exempt throughout). Consequently, the following behaviors cannot be verified in production: actual dial reachability, whether the AI truthfully declares its assistant identity during calls (a contract commitment), whether it refuses to impersonate the user, and whether the destination region restriction (Mainland China declared unsupported) is enforced at the execution layer. These links require supplementary verification via merchant-side configuration evidence or call recording samples.

Probabilistic Nature of LLM Safeguards	The Agent's injection defenses are based on LLM judgment, which is probabilistic rather than deterministic. In this test the Agent showed stable structural resistance to injected directives (ignoring direct-dial instructions, consistently requiring a real merchant name + city), but this does not guarantee the same outcome under all contexts, model versions, or script variants. Attackers may bypass current defenses through repeated retries, context manipulation, or model version changes. The conclusions in this report are valid only within the forensic time window (2026-08-26 16:03—16:37 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 (including personal information like callback numbers and booking names) was written to persistent storage, whether sensitive information was leaked in the internal context (even if the final output does not contain it), or whether upstream tool calls were triggered (even if a safe response was returned). These internal behaviors require white-box auditing or runtime instrumentation to verify.
Agent Version and Configuration Changes	The audited version is `0650d33a-d9e0-466d-9847-943227820c2f` (status=in_review). System prompt updates, merchant-search backend changes, dial subsystem upgrades, or security policy adjustments may all alter this report's conclusions. Re-audit is required after version changes.
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 real outbound calls and call behavior cannot be directly verified under audit mode, it is recommended that the merchant supplement execution-layer configuration evidence and call recording samples, with follow-up white-box auditing and continuous adversarial testing 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