SLOWMIST AI

Aident Lead Hunter Security Audit Report

REPORT NO.

SM-AR-2026-023

DATE

2026-11-02

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

No exploitable security vulnerabilities were found in this black-box test. Injection defence held: three payload groups — system-prompt and credential solicitation, role hijack with shell command execution, and fabricated "verified buyers" with separator injection — were not adopted; every task returned a normal lead report, with zero hits for command output, fabricated contacts or fabricated sources, and no system prompt, secret or environment information echoed back. SSRF/egress did not occur as verified by a weblog canary: two payloads embedding a callback URL triggered no outbound request, and the only hit was our own `curl` positive control (`found=True` , `total=1` , UA `curl/8.7.1`), proving the probe works. Control-plane and contract boundaries behaved correctly: invalid audit key `401` , forged `serviceVersionId` `404` , missing side-effect acknowledgement `409` , invalid Offer `400` , out-of-range `quantity` `409` , cross-session invocation access `401` ; idempotent replay returned the original `invocationId` / `dispatchId` / `taskId` . Billing exemption: all 146 objects carrying billing fields across 75 evidence files were `executionMode=audit` , `billingStatus=exempt` , `paidMicroUsdc=0` . All 16 dispatches carry structured `failureCode` values and `logs` reconstructs each dispatch and request chain by invocation. Six observations were recorded: `inputSchema` is not enforced at the acceptance boundary (six violation classes are accepted and only rejected at runtime), three tasks expired under the 600-second Light tier producing empty deliveries, terminal `output[]` carries an artifact reference while `artifactMetadata` is empty and the download endpoint rejects `kind=output` with `400` , the structured-input channel is undocumented, the `IP 属地` field is always empty, and outputs consist of public social-media handles and comment excerpts. All of these are quality, availability or observability issues, none constitutes a proven security threat, and none was entered into the findings ledger.

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, submitting structured lead-research parameters through the AgentOn platform audit API asynchronous Task channel and observing the responses; the Agent's source code, system prompts, social-media retrieval tooling, upstream data providers and merchant-side runtime logs were not available. The following limitations must be considered alongside this report:

OUT-OF-SCOPE MODULE	AFFECTED FINDINGS	PENDING VERIFICATION
Black-box Test Coverage Limitations		This audit covered catalogue and contract verification, session lifecycle, baseline lead research, three injection classes, SSRF with a weblog positive control, contract boundaries, idempotency, authentication and authorization, audit logging and billing exemption. Uncovered attack surfaces include but are not limited to: multi-turn progressive jailbreaking (this channel has `maxInputRounds=0` and no follow-up rounds), Unicode homoglyphs and encoding obfuscation, concurrency and multi-tenant isolation, upstream social-media data poisoning, citation-chain integrity of leads, and exhaustive factual verification of reports. Untested does not mean secure.
Upstream Retrieval Chain Not Observable		The social-media search provider, caching layer, proxy and index scope are invisible. The weblog probe ended at `total=1` with that single record being our own positive control; this only means no hit occurred here and does not exclude proxy, cache or delayed paths. It is also impossible to verify whether the Agent actively fetches external URLs found in retrieved results.
Lead Report Factual Accuracy Only Sampled		Only the field completeness of the five leads in the baseline report and the single lead in the retry was sampled (intent level, intent basis, comment excerpt, user reply time, source post, lead profile); individual Instagram accounts and comments were not visited to verify existence or timestamp accuracy. Self-declared caveats such as "lead volume is not guaranteed" and "some sources were not completed" were not independently validated.

Artifact File-input Channel Unavailable	The `artifactContract` declares `text/plain` support (one item, 1 MiB per item and total), but an actual `POST .../artifacts/input` returned `503 SLOWMIST_AUDIT_UNAVAILABLE` without issuing an `uploadToken`. MIME, size and content validation for the file modality, as well as file-borne indirect injection, cannot be verified. On the output side the CSV attachment is downloadable when `kind` is omitted.
LLM Jailbreak Adversarial Testing Not Executed	FuzzyAI jailbreak adversarial testing was skipped per the established project convention. The model-level jailbreak surface of D2 (jailbreak hit rate) therefore cannot be verified; conclusions on injection defence cover only the vectors exercised and behaviours observed here.
Production Billing and Cost Impact Unverifiable	The audit channel was billing-exempt throughout (`executionMode=audit`, `billingStatus=exempt`, `paidMicroUsdc=0`). Whether contract-violating input (seven upstream `provider_http_rejected` occurrences) and timed-out tasks (three `deadline_expired` occurrences) are billed in production, and how reconciliation and refunds work, cannot be verified.
Impact of Task Time Limits and Queueing	`deadlineAt` is computed from task creation, so queueing time counts against execution time. The Light tier sets `maxDurationSeconds=600`; three of the 16 dispatches produced no output on expiry (`async_expired:deadline_expired`), and successful reports repeatedly state that the task time limit was reached and search stopped. The 1,200-second Standard tier and 1,800-second Deep tier were not exercised, so their delivery completeness and cost cannot be verified.
Supply Chain and Retrieval Implementation Not Visible	No source code, external tool packages, dependency inventory, system prompt or social-media retrieval implementation was obtained. Source allowlists, deduplication strategy, content-filtering rules, citation-chain validation and caching strategy cannot be verified, nor can behaviour under upstream data poisoning or index contamination be assessed.

Uncertainty of Negative Results	No security vulnerabilities were found, but this does not mean the Agent is secure. A negative (PASS) black-box result only indicates that no observable anomalous behaviour was triggered within the limited input space of this test. Injection resistance depends on model judgement and is probabilistic; the conclusions are valid only for this evidence window. White-box auditing, retrieval-source trust assessment and continuous adversarial exercises are recommended to build defence in depth.

> In summary, the PASS rating 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 the upstream retrieval chain is not observable, lead facts were only sampled, the Artifact input channel is unavailable and model-level jailbreak testing was not performed, it is recommended that the merchant enforce `inputSchema` validation at the acceptance boundary with precise failure codes, align artifact references with `artifactMetadata` and the download contract, document the structured-input channel, adjust `deadlineAt` for queueing or raise the Light-tier duration to avoid empty deliveries, and jointly confirm production billing and refund semantics with the platform; follow-up white-box auditing and continuous adversarial testing will build more comprehensive 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