SLOWMIST AI

LittlePinkPotato Security Audit Report

REPORT NO.	DATE
SM-AR-2026-006	2026-08-06

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

Of 30 applicable checks, 28 passed, 2 NA adjustments, no CRITICAL/HIGH/MEDIUM/LOW/INFO security findings. This audit found no security vulnerabilities within the black-box testing scope. Offer draft status and input_contract declaration completeness issues are contract/availability matters, not counted in the security rating, and are declared in scope limitations.

Severity Distribution

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

Scope & Limitations

This audit is a third-party black-box test of LittlePinkPotato via the AgentOn platform audit API, with no source code, no MCP/Skill packages, non-Web3, single-agent. Per user request, no jailbreak adversarial testing was performed. 44 checks marked NA (not applicable) and excluded from scoring. This audit found no security vulnerabilities, but the following contract and availability matters must be considered alongside this report:

OUT-OF-SCOPE MODULE	AFFECTED FINDINGS	PENDING VERIFICATION
Black-box test coverage limitation		This audit only sends requests via the AgentOn audit API and observes response behavior; it cannot access Agent source code, configuration, system prompts, internal tool implementations, or runtime logs. 44 of 74 baseline checks marked NA due to black-box/no-source/non-Web3/single-agent; the security rating is based on 30 applicable checks. Uncovered attack surfaces include but are not limited to: multi-turn conversation injection, encoding bypass, indirect injection via file upload, cross-session data leakage, supply chain attacks. Not tested does not mean secure.
Offer draft status and actual availability		All Offers have `draft` status, but agent top-level `audit_eligible=true`, `contract_test_passed=true`. Contract tests pass via probes carrying `contract_test=true` to bypass draft restrictions; audit requests without this flag trigger actual upstream handling logic, and actual availability may be inconsistent with the `audit_eligible` declaration. After Offers transition to `published`, complete P1/P2 acceptance must be re-executed.
Incomplete input_contract declarations		`input_contract` declares `modalities: [text, image, video, file]` without language requirements, required fields, or input format constraints. Upstream applies input validation not declared in the contract; auditors/integrators cannot infer valid input format from `input_contract`. This is a contract consistency issue, not counted in the security rating.
Jailbreak adversarial testing not performed		Per user request, no jailbreak adversarial testing was performed (D2-S2 runtime jailbreak verification marked NA). Agent LLM security protection is probabilistic; not performing jailbreak testing does not mean the Agent can resist all jailbreak attacks. Future evaluation of whether to include jailbreak adversarial testing is recommended.

Weblog exfiltration test limited	This weblog exfiltration test (D4 SSRF/data exfiltration verification) returned `found=False`, limited by Offer draft status causing some normal paths unavailable; external connection testing could not be executed on the normal path for all Offers. This result is not sufficient evidence of outbound blocking; it only indicates no outbound connection was observed this time. Supplemental outbound testing on the normal execution path is needed after all Offers transition to `published`.
Probabilistic nature of LLM protection	Agent security protection is based on LLM judgment and is probabilistic, not deterministic. The same input may produce different responses at different times, in different contexts, or under different model versions. Black-box testing cannot enumerate all conditions that trigger malicious outputs. This report's conclusions are valid only within the evidence-collection window (2026-08-06 15:19—15:35 UTC+8).
Unobservable internal behavior	Black-box testing can only observe final outputs and cannot detect whether unexpected internal behavior occurred in the Agent, for example: whether partial operations were executed before rejection, whether user input was written to persistent storage, whether sensitive information was leaked in internal context, whether upstream tool calls were triggered. These internal behaviors require white-box audit or runtime instrumentation to verify.
Agent version and configuration changes	Agent version and configuration may change after the audit. System prompt updates, toolchain changes, upstream model upgrades, or security policy adjustments may introduce new risks. This report's conclusions do not apply to the Agent after version changes; a re-audit is required after changes. All Offers are currently `draft`; after transition to `published`, P1/P2 acceptance must be re-executed.
Uncertainty of negative results	The PASS rating of this audit should be understood as "no CRITICAL/HIGH/MEDIUM/LOW/INFO security vulnerabilities found within the limited input space of this black-box test," not as "the Agent has been proven secure." Security is a continuous process; regular regression audits are recommended, combined with white-box audit and runtime monitoring to build a defense-in-depth verification system.

In summary, this audit found no security vulnerabilities within the black-box testing scope, rated PASS. However, contract and availability issues such as Offer draft status and incomplete input_contract declarations must be resolved before production. After Offers transition to `published`, complete P1/P2 acceptance must be re-executed. Future supplementation with white-box source audit, post-publish normal-path verification, jailbreak adversarial testing, and continuous runtime monitoring is recommended.

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