SLOWMIST AI

MistTrack Agent Security Audit Report

REPORT NO.	DATE
SM-AR-2026-007	2026-08-14

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

White-box source code audit + black-box test verification. No CRITICAL/HIGH/MEDIUM/LOW security vulnerabilities identified. The v2 code fixes all 4 LOW findings from the v1 audit: exception information leakage, missing injection defense in system prompt, idempotency TOCTOU race condition, and absence of structured logging. New security measures include code-level extraction validation, address format regex, centralized error handling, and MistTrack error isolation. HMAC-SHA256 signature verification + nonce replay prevention + timestamp skew detection all verified as passing.

Severity Distribution

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

Scope & Limitations

This audit is a third-party white-box source code audit + black-box test verification of MistTrack Agent v2, conducted through source code review and the AgentOn platform audit API. The audit covers 12 Python files (`main.py` , `security.py` , `nlu.py` , `offers.py` , `storage.py` , `misttrack_client.py` , `config.py` , `errors.py` , `context_probe.py` , `formatting.py` , `logging_utils.py` , `__init__.py`) + Dockerfile + docker-compose.yml + PREPARE.md + pyproject.toml + uv.lock + smoke test scripts. No jailbreak adversarial testing was performed. The following caveats must be considered when reading this report:

OUT-OF-SCOPE MODULE	AFFECTED FINDINGS	PENDING VERIFICATION
White-Box Audit Coverage Limitations		This audit is based on the source code snapshot as of 2026-08-14, covering 12 Python files and configuration files. The audit focused on authentication signatures, prompt injection defense, SSRF/data exfiltration, idempotency mechanisms, error handling, and logging audit dimensions. Uncovered attack surfaces include but are not limited to: multi-turn encoding bypass injection, indirect injection via file upload, cross-session data leakage, supply chain attacks, DoS attacks, and SQLite concurrent write races. Untested does not mean secure.
Jailbreak Adversarial Testing Not Performed		FuzzyAI automated jailbreak adversarial testing was not performed in this audit. The agent's LLM-based security protection is probabilistic. The v2 code adds injection defense rules (Rule 6) and `<user_message>` tag wrapping at the system prompt level, but absence of jailbreak testing does not imply the agent can withstand all jailbreak attacks. It is recommended to evaluate whether to include jailbreak adversarial testing in future assessments.
Probabilistic Nature of LLM Protection		The agent's security protection is based on LLM judgment, which is probabilistic rather than deterministic. The v2 code adds code-level extraction validation (`_validate_extraction`) as a deterministic supplement to LLM judgment — address/txid must appear verbatim in user-authored text and pass format regex — but the LLM summarization layer remains probabilistic. The same input may produce different responses at different times, in different contexts, or under different model versions. The conclusions in this report are valid only within the evidence collection window.

Agent Version and Configuration Changes	This audit is based on the v2 source code snapshot. 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.
Uncertainty of Negative Results	The PASS rating in this audit should be understood as "no CRITICAL/HIGH/MEDIUM/LOW level security vulnerabilities were found within the scope of this white-box source code audit and black-box testing," rather than "the agent has been proven secure." Security is a continuous process. Regular regression audits are recommended, combined with jailbreak adversarial testing and runtime monitoring to build a defense-in-depth verification system.

> In summary, no security vulnerabilities were identified in this white-box source code audit + black-box test verification, and the rating is PASS. The v2 code fixes all 4 LOW findings from the v1 audit and introduces new security measures including code-level extraction validation, address format regex, centralized error handling, MistTrack error isolation, and structured logging. HMAC-SHA256 signature verification, nonce replay prevention, timestamp skew detection, and atomic idempotency operations were all verified as passing. It is recommended to supplement with jailbreak adversarial testing and continuous runtime monitoring in future assessments.

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