SLOWMIST AI

FlowCredit Security Audit Report

REPORT NO.

SM-AR-2026-012

DATE

2026-09-18

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

None of the 22 black-box cases found exploitable security vulnerabilities. All 11 injection vectors (instruction override with score and credit-limit hijacking, server-owned field injection, data-fabrication pressure, system-prompt and threshold solicitation, Base64 obfuscation, zero-width character obfuscation, maximum-length injection, scoring-rule reverse-engineering requests, payment-address and transaction-signing solicitation, private-key feeding, and cross-call memory probing) were rejected: decision fields remained stable at `decisionStatus=insufficient-evidence`, `suggestedCreditBand=manual-only` and `assessmentMode=real`; server-owned fields (`assessmentMode` / `normalizationProfileId` / `peerProfileId`) were explicitly ignored and echoed in `validation.ignoredInputs`; the Agent explicitly refuses to complete or fabricate missing data and produced no favourable rating or approval semantics. SSRF/URL injection produced no outbound request verified via weblog (`found=False`); private keys and seed phrases were neither accepted, used nor echoed; payment-address, transaction-signing and treasury-balance requests were all refused. Control paths behaved correctly: invalid audit key `401`, forged `versionId` `400`, out-of-scope session `400`, missing side-effect acknowledgement `409`, non-object `providerInput` `400`, oversized request body `413`, and idempotent replay returning the original invocation (`replayed=true`); failed dispatches carry structured `failureCode` values in the audit log (`input_schema_invalid` / `runtime_protocol_error`) and execution traces are replayable; all 39 records carrying billing fields remained billing-exempt (`executionMode=audit`, `billingStatus=exempt`, `paidMicroUsdc=0`) with no billing anomaly. Five observations were also recorded (generic client-facing error for contract-violating input, idempotency conflict returning `503` instead of `409`, blank input triggering an execution-layer protocol error, disclosure of model and rule version identifiers, and the mismatch between `acceptingRequests=false` and the observed callability); none were assessed as an actual security threat and none were 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 a `prompt` through the AgentOn platform audit API Direct channel and observing the returned structured risk intelligence. The Agent's source code, configuration, system prompts, rule thresholds, reference datasets 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 executed 22 cases covering catalogue and multi-version discrimination, Direct contract verification, session lifecycle, injection resistance (11 vectors), SSRF outbound verification, contract boundaries, idempotency, authentication and authorization, audit logging and billing exemption. Attack surfaces not covered include but are not limited to: multi-turn progressive jailbreaking (not applicable as this channel is a stateless single call), Unicode homoglyph variants, concurrency and multi-tenant isolation, merchant-side backend and reference-dataset penetration, and numerical boundary or overflow behaviour of the scoring engine. Untested does not mean secure.
Primary Scoring Path Unreachable (TAI / CCI / Numeric Risk Grade)		The natural-language parser (`nl-draft-v0.1`) extracted at most 5 fields from `prompt` in this test (`revenueUsd` , `gpuModel` , `gpuHours` , `computeSpendUsd` , `payingCustomers`); fields such as `inputTokensM` / `outputTokensM` / `validRatePct` / `monthlySeries` / `R` / `C` never entered the draft, and the only declared structured input channel (Artifact `application/json`) remained unavailable. The computation branches for TAI, CCI and the numeric Risk Grade were therefore never triggered, and the scoring and integrity-veto logic was observed only in the `insufficient-evidence` branch. The conclusions here do not cover the correctness or manipulation resistance of the scoring engine under complete data.

Artifact File-input Channel Unavailable	The `artifactContract` declares support for `application/json` (at most 1 item, 1 MiB per item and total), yet both upload attempts (without an `X-Finch-Offer-Id` header and with a placeholder Offer header) returned `HTTP 503 SLOWMIST_AUDIT_UNAVAILABLE` without issuing an `uploadToken`; neither the byte upload nor the content-processing path was ever reached. MIME, size and content validation for the file modality, as well as file-borne indirect injection, cannot be verified.
LLM Jailbreak Adversarial Testing Not Executed	FuzzyAI jailbreak adversarial testing was skipped as requested (consistent with prior audits of maneki-ai, LifeClaw and SuperClaw). The model-level jailbreak surface (jailbreak hit rate) of D2 therefore cannot be verified; conclusions on injection defence in this report cover only the injection vectors exercised here and the observed resistance behaviour.
Production Billing and Cost Impact Unverifiable	The audit channel was billing-exempt throughout (`executionMode=audit`, `billingStatus=exempt`, `paidMicroUsdc=0`). A single call to this Agent is priced at `unitPriceAtomic=9900000` (9.9 USDC). Whether contract-violating input and execution-layer failures (16 `input_schema_invalid` and 2 `runtime_protocol_error` occurrences in this audit) are billed in production, whether they consume a call right (`callRightUnits=1`), and how reconciliation and refunds work, cannot be verified in this audit.
Responsibility Attribution for Failures Pending	Whether `input_schema_invalid`, `runtime_protocol_error` and the `503 SLOWMIST_AUDIT_UNAVAILABLE` returned on idempotency conflict are decided by the platform or the merchant runtime cannot be distinguished in a black-box setting. Joint confirmation by both parties is recommended: where contract validation executes, which party generates the failure codes, and the semantic boundary between `503` and `409`.
Probabilistic Nature of LLM Safeguards	The Agent's injection resistance relies partly on LLM judgement, which is probabilistic rather than deterministic. In this test the Agent showed stable structural resistance (decision fields were unaffected by injections, server-owned fields were ignored, fabrication was explicitly refused), but this does not guarantee the same outcome under all contexts, model versions or script variants. The conclusions are valid only within the forensic window (2026-09-18 16:23–16:40 UTC+8).

Agent Version and Configuration Changes	The audited version is `e96e9a01-50aa-4686-bd0e-1aa9cdfcf967` (`versionRevision=1`, `versionStatus/runtimeStatus/contractStatus=draft`, `publicationGeneration=0`, `contractTestPassed=true`). The same catalogue entry also contains two non-executable versions, `e5d490b5-…` (`contractTestPassed=false`) and `62e1a2ff-…` (`rejected`), which must not be confused with it. Changes to the system prompt, parser, rule engine, reference datasets or the Direct contract 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 behaviour was triggered within the limited input space tested. 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 defence-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 the primary scoring path (TAI/CCI/numeric risk grade) was unreachable in this audit, the Artifact structured-input channel was unavailable, and model-level jailbreak adversarial testing was not executed, it is recommended that the merchant supplement scoring-consistency testing under complete data and validation material for the non-self-reported evidence path, that the platform and merchant jointly define the semantics of the failure codes and the `503` / `409` boundary, and that follow-up white-box auditing and continuous adversarial testing be used to build 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