SHADOW AI
DISCOVERY · SECSTUDIO

PREPARED FOR CLIENT LOGO

DETAILED FINDINGS REPORT

AI Security Findings — Technical Detail

Prepared for **Northwind Trading Co.**

REPORTING PERIOD	SCOPE	CLASSIFICATION
Q1 2026 · as of 12 Aug 2026	**320 endpoints · fleet-wide**	**Confidential**
PREPARED BY	REPORT ID	FINDINGS
Shadow AI Discovery	**SAD-FIND-2026-0142**	**24 grouped · 39 instances**

CONFIDENTIAL — prepared for the named recipient. Do not distribute without authorization.

SECTION 1

Summary of findings

24 distinct issues were identified across 320 endpoints, affecting 39 asset instances. Findings are grouped by issue type — each entry below lists every affected host, user and path. Every finding carries an **AI confidence score** from the reasoning engine, which validates the detection against its evidence to suppress false positives.

4	13	12	7	3
CRITICAL	HIGH	MEDIUM	LOW	INFO

FINDINGS REGISTER (GROUPED)

ID	SEVERITY	FINDING	CATEGORY	AFFECTED	AI CONF.	STATUS
F-001	Critical	**API keys pasted into public LLMs**	Secret exposure	14	96%	Open
F-002	Critical	**Unregistered MCP server (vulnerable)**	Vulnerable component	6	93%	Open
F-003	High	**Unapproved AI coding agents**	Unsanctioned tool	31	89%	In progress
F-004	High	**Typosquatted MCP package "server-filesysem"**	Typosquat	3	91%	Open
F-005	High	**Offboarded-user AI access still active**	Offboarded access	9	85%	Open
F-006	High	**LLM SDK on vulnerable version**	Vulnerable component	22	88%	Open
F-007	Medium	**Shadow browser AI extensions**	Browser extension	48	78%	Open
F-008	Medium	**Local model with unmonitored network egress**	Unsanctioned egress	5	74%	Open
F-009	Medium	**MCP server with over-broad tool scope**	Agent blast-radius	4	81%	Open
F-010	Low	**Unencrypted (http://) calls to LLM endpoint**	Transport	7	69%	Open

Showing the 10 highest-priority groups; F-011 – F-024 (medium/low/info) follow the same structure and are included in the machine-readable export.

CRITICAL FINDINGS

Critical · AI confidence 96% · F-001 · Secret exposure · OWASP LLM06 · CWE-522

API keys pasted into public LLM prompts

First seen **04 Aug 2026** Instances **14** Priority **P1**

DESCRIPTION

Provider API keys (OpenAI, Anthropic, internal) were observed pasted into public LLM chat prompts. Keys sent to a third-party service must be treated as compromised — they can be replayed to run up cost, exfiltrate data, or pivot into connected systems.

AFFECTED ASSETS (14 — FIRST 6 SHOWN)

HOSTNAME	IP ADDRESS	USER	PATH / EVIDENCE
NW-LT-4821	10.2.14.8	j.rivera	`chrome://chatgpt · prompt body · sk-…REDACTED`
NW-LT-3390	10.2.11.51	a.okafor	`~/.config/agent/history.json:42`
NW-WS-1177	10.4.6.23	s.klein	`ChatGPT desktop · pasted ANTHROPIC_API_KEY`
NW-LT-2043	10.2.14.61	d.mehta	`~/projects/etl/.env → prompt`
NW-LT-5510	10.2.19.7	l.zhang	`clipboard → claude.ai · sk-…REDACTED`
NW-WS-0932	10.4.6.88	t.novak	`~/.zsh_history:2210`

+ 8 more instances (users p.singh, m.ali, r.costa, …). Full list in the JSON export.

IMPACT

Immediate financial and data-exfiltration risk; keys with broad scope can reach production data stores and CI. Two keys were still live at time of report.

REMEDIATION

Rotate all 14 keys now; scope replacements to least privilege; enable inline AI-DLP to block secret egress at the endpoint; brief the affected users.

Critical · AI confidence 93% · F-002 · Vulnerable component · SHAI-2026-001 · CWE-1104

Unregistered MCP server on a vulnerable version

First seen **02 Aug 2026** Instances **6** Priority **P1**

DESCRIPTION

An MCP (Model Context Protocol) server was found running unregistered on a version affected by advisory SHAI-2026-001 (tool-poisoning / path traversal). It exposes filesystem and shell tools to any connected agent.

AFFECTED ASSETS (6)

HOSTNAME	IP ADDRESS	USER	PATH / EVIDENCE
NW-LT-4821	10.2.14.8	j.rivera	`mcp://localhost:8931 · v0.3.1 · /opt/mcp/fs`
NW-LT-3390	10.2.11.51	a.okafor	`~/.mcp/servers.json:12`
NW-DEV-0210	10.6.2.4	build-svc	`/etc/mcp/fs-server.yaml`
NW-LT-2043	10.2.14.61	d.mehta	`mcp://localhost:8931 · v0.3.1`
NW-LT-1180	10.2.10.9	k.brandt	`~/.mcp/servers.json:8`
NW-DEV-0212	10.6.2.6	build-svc	`/etc/mcp/fs-server.yaml`

EVIDENCE

```
$ mcp inspect localhost:8931
server: filesystem  version: 0.3.1  (advisory SHAI-2026-001)
tools: read_file, write_file, run_shell   scope: /  (unrestricted)
```

REMEDIATION

the 6 endpoints; upgrade past the fixed version; restrict tool scope; register the server in the approved inventory or remove it.

HIGH FINDINGS

High · AI confidence 89% · F-003 · Unsanctioned tool · OWASP LLM · CWE-829

Unapproved AI coding agents with code & credential access

First seen **28 Jul 2026** Instances **31** Priority **P2**

Coding agents not on the approved list were found with access to source repositories and local credentials. They can read code, run commands, and reach connected services outside policy.

AFFECTED ASSETS (31 — FIRST 5 SHOWN)

HOSTNAME	IP ADDRESS	USER	PATH / EVIDENCE
NW-LT-2043	10.2.14.61	d.mehta	~/.local/bin/aider · repo: payments-api
NW-DEV-0210	10.6.2.4	build-svc	/opt/agent/ · CI runner token in env
NW-LT-5510	10.2.19.7	l.zhang	VS Code ext · workspace: core-ledger
NW-LT-1180	10.2.10.9	k.brandt	~/.config/agent/ · GH_TOKEN present
NW-LT-3390	10.2.11.51	a.okafor	~/dev/tools/agent-cli

+ 26 more instances across engineering. Full list in the JSON export.

REMEDIATION

Approve one sanctioned agent, block the rest by policy, route exceptions through review, and scope tokens to least privilege.

High · AI confidence 85% · F-005 · Offboarded access · CWE-284

Offboarded users still holding AI-tool access

First seen **19 Jul 2026** Instances **9 identities** Priority **P2**

Nine departed users retain active access to AI tools and/or live API keys — access that should have been revoked at offboarding.

AFFECTED IDENTITIES (9 — FIRST 5 SHOWN)

USER	LEFT	LAST HOST	RESIDUAL ACCESS
c.duarte	12 Jun 2026	NW-LT-7701	OpenAI key active · agent config present
m.ali	28 May 2026	NW-LT-6620	ChatGPT Team seat active
r.costa	02 Jun 2026	NW-WS-0410	Anthropic key active
p.singh	15 May 2026	NW-LT-5309	MCP token valid
e.haas	30 Jun 2026	NW-LT-4102	Copilot seat active

+ 4 more. Wire directory sync (SCIM) so AI-tool access is revoked automatically at offboarding.

AI CONFIDENCE SCORING

Every finding is scored by the reasoning engine, which reads the raw detection and its evidence and judges whether the issue is real. The **AI confidence score** (0–100%) reflects how strongly the evidence supports the finding; low-confidence detections are flagged or auto-classified as false positives so analysts spend time only on real issues.

HOW THE SCORE IS BUILT

Evidence strength	Does the artifact/log directly prove it?
Source authority	Deterministic detector vs heuristic
Corroboration	Multiple signals across the fleet
Context	Sanctioned/approved-list checks

READING THE SCORE

85–100%	High confidence — act now
60–84%	Likely — verify then act
< 60%	Review — may be a false positive

SEVERITY MODEL

SEVERITY	MEANING	TARGET SLA
Critical	Active exposure or exploitable path to sensitive data	7 days
High	Significant risk; likely exploitable or policy breach	30 days
Medium	Meaningful risk requiring remediation	90 days
Low	Minor / hardening	180 days

SCOPE & METHODOLOGY

Endpoint sensors discover AI usage across the fleet (9 collectors: API keys, CLI tools, MCP configs, LLM SDKs, extensions, installed apps, network, DNS, agent skills). Detections are de-duplicated and grouped by issue type; each group lists every affected host, user and path. The AI reasoning layer validates and scores each finding to reduce false positives. This report pairs with the Executive Summary (SAD-EXEC-2026-0142) and the Fleet/AIBOM report (SAD-FLEET-2026-0142).