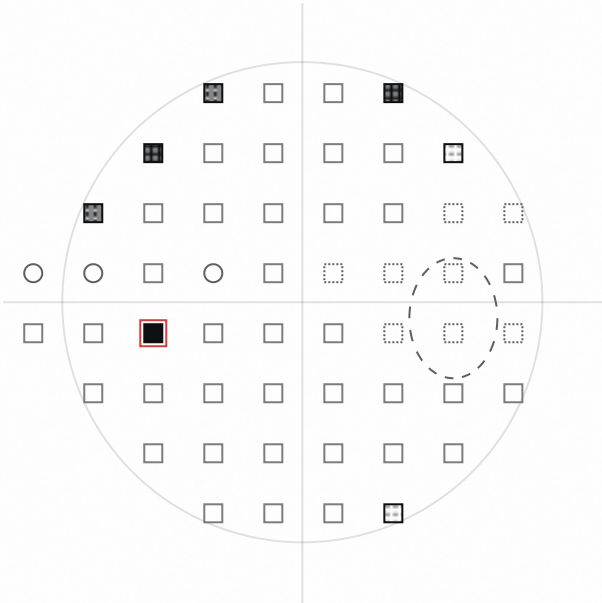


SAMPLE-01 / VERSION 1.1, AFTER THE RETEST

Penetration test report

AI Pentest and Stack Pentest, one scope: Launch Clearance

CLIENT	
SYSTEM	Claims assistant (an AI agent with tools) and the claims API behind it
TEST WINDOW	2026-08-17 to 2026-08-28
REPORT	2026-09-03, retest 2026-09-24
TESTED BY	Scotoma
DISTRIBUTION	Client security team



54 CATEGORIES / 7 FINDINGS /
3 N/A / 8 OUT OF SCOPE

CRITICAL HIGH MEDIUM LOW
 TESTED, NOTHING FOUND NOT APPLICABLE
 OUT OF SCOPE

The only blind spot in our report is the
one you scope out.

1 / Executive summary

We tested the claims assistant on the client’s customer portal and the claims API behind it, in staging, from 17 to 28 August 2026. We found seven issues: one critical, two high, two medium and two low.

The critical one is a chain. A knowledge-base page that nobody reviewed could give the assistant instructions, the assistant could call its refund tool outside the claim it was handling, and the claims API trusted the assistant’s token for every tenant. Together they let anyone with a partner upload account read and change another customer’s claim, without opening the chat themselves.

At the retest on 24 September 2026, five findings were fixed, one was partly fixed, and the client accepted the risk of one.

SEVERITY		
	CRITICAL	1
	HIGH	2
	MEDIUM	2
	LOW	2

FIX FIRST	
F-01: scope the assistant’s token to one tenant and enforce claim ownership.	
F-03: keep retrieved text out of the instruction channel.	

2 / Scope and dates

SYSTEMS	The claims assistant (web chat, English and Dutch), its three tools (claim lookup, document upload, refund request), the knowledge base it retrieves from, the claims API (REST, 23 endpoints) and the address-lookup MCP server, with the vendor’s written permission. In the cloud account, only what the assistant touches: the claim document storage, the secrets it uses, its hosted model’s keys and quotas, and the audit trail of its tool calls.
SCOPED OUT	The rest of the cloud account: identities and roles, privilege paths, sign-in, pipelines, instance metadata, containers, serverless functions and external infrastructure. That is a Cloud & Infrastructure test, not part of Launch Clearance. They are dashed on the cover: not tested, and the report says so.
NOT APPLICABLE	Three agentic categories describe parts this system does not have: code execution, messages between agents, and failures that spread across several agents. The matrix gives the reason for each.
ENVIRONMENT	Staging, configured like production, with synthetic customers and claims.
ACCESS	Grey box: the system prompt, the tool definitions and the API documentation; two test tenants with two roles each.
TEST WINDOW	2026-08-17 to 2026-08-28, 10 testing days. Rules of engagement signed 2026-08-12.
REPORT	Version 1.0 on 2026-09-03; this version 1.1 after the retest on 2026-09-24.

3 / Method and coverage

Each category was tested, marked not applicable with the reason, or scoped out. Findings carry their id; a category marked tested was tested and nothing was found. AI findings were reproduced ten times each, with the model, version and settings recorded.

OWASP TOP 10 FOR LLM APPLICATIONS 2026			OWASP TOP 10 FOR AGENTIC APPLICATIONS 2026		
ID	CATEGORY	RESULT	ID	CATEGORY	RESULT
LLM01:2026	Prompt Injection	F-03 / HIGH	ASI01	Agent Goal Hijack	TESTED
LLM02:2026	Sensitive Information Disclosure	TESTED	ASI02	Tool Misuse and Exploitation	F-02 / HIGH
LLM03:2026	Excessive Agency	TESTED	ASI03	Identity and Privilege Abuse	TESTED
LLM04:2026	Supply Chain	TESTED	ASI04	Agentic Supply Chain Vulnerabilities	F-05 / MEDIUM
LLM05:2026	Data and Model Poisoning	TESTED Retrieval data tested; training and fine-tuning not applicable (hosted model).	ASI05	Unexpected Code Execution (RCE)	NOT APPLICABLE The assistant has no tool that runs code.
LLM06:2026	Unbounded Consumption	TESTED	ASI06	Memory & Context Poisoning	TESTED
LLM07:2026	Misinformation	TESTED	ASI07	Insecure Inter-Agent Communication	NOT APPLICABLE One agent, so no messages pass between agents.
LLM08:2026	Hidden Context Exposure	F-04 / MEDIUM	ASI08	Cascading Failures	NOT APPLICABLE One agent in one workflow, so a failure has no other agent to spread to.
LLM09:2026	Vector and Embedding Weaknesses	TESTED	ASI09	Human-Agent Trust Exploitation	TESTED
LLM10:2026	Improper Output Handling	TESTED	ASI10	Rogue Agents	TESTED

3 / Method and coverage, continued

OWASP API SECURITY TOP 10 (2023)

ID	CATEGORY	RESULT
API1:2023	Broken Object Level Authorization	F-01 / CRITICAL
API2:2023	Broken Authentication	TESTED
API3:2023	Broken Object Property Level Authorization	TESTED
API4:2023	Unrestricted Resource Consumption	TESTED
API5:2023	Broken Function Level Authorization	TESTED
API6:2023	Unrestricted Access to Sensitive Business Flows	TESTED
API7:2023	Server Side Request Forgery	TESTED
API8:2023	Security Misconfiguration	TESTED
API9:2023	Improper Inventory Management	TESTED
API10:2023	Unsafe Consumption of APIs	TESTED

CLOUD AND INFRASTRUCTURE

CATEGORY	RESULT
Least privilege for users, roles and service accounts	OUT OF SCOPE
Privilege escalation paths and cross-account trust	OUT OF SCOPE
Sign-in hardening: federation, MFA and conditional access	OUT OF SCOPE
Object storage exposure and access policies	TESTED
Secrets in code, images and secret stores	TESTED
CI/CD pipeline integrity and build permissions	OUT OF SCOPE
Instance metadata and workload identity protection	OUT OF SCOPE
Container and Kubernetes configuration	OUT OF SCOPE
Serverless functions and event triggers	OUT OF SCOPE
Hosted AI services: model endpoints, keys and quotas	TESTED
Logging and audit trail coverage	F-07 / LOW
External infrastructure: exposed hosts, services and remote access	OUT OF SCOPE

OWASP WSTG CATEGORIES

ID	CATEGORY	RESULT
WSTG-INFO	Information Gathering	TESTED
WSTG-CONF	Configuration and Deployment Management Testing	TESTED
WSTG-IDNT	Identity Management Testing	TESTED
WSTG-ATHN	Authentication Testing	TESTED
WSTG-ATHZ	Authorization Testing	TESTED
WSTG-SESS	Session Management Testing	F-06 / LOW
WSTG-INPV	Input Validation Testing	TESTED
WSTG-ERRH	Testing for Error Handling	TESTED
WSTG-CRYP	Testing for Weak Cryptography	TESTED
WSTG-BUSL	Business Logic Testing	TESTED
WSTG-CLNT	Client-side Testing	TESTED
WSTG-APIT	API Testing	TESTED

4 / Findings

F-01

CRITICAL

CVSS-B 9.3

API1:2023

STACK LAYER

The claims API serves any tenant’s claim to the assistant’s token

WHAT WE FOUND

The assistant calls the claims API with one service token. The API checked that the token was valid, not that the claim belonged to the customer in the conversation, so a claim of another tenant was returned, and the update endpoint accepted changes to it.

BUSINESS IMPACT

Anyone who can steer the assistant can read and change another customer’s claim.

EVIDENCE

Requests REQ-07 to REQ-11, appendix B

FIX

Check authorization on every object, not only at login: scope the assistant’s token to one tenant and enforce claim ownership in the API.

RETEST

FIXED 2026-09-24

CVSS:4.0/AV:N/AC:L/AT:N/PR:N/UI:N/VC:H/VI:H/VA:N/SC:N/SI:N/SA:N

F-02

HIGH

CVSS-B 8.2

ASI02

MODEL LAYER

The assistant can call its refund tool for a claim it is not handling

WHAT WE FOUND

The refund tool takes a claim reference as an argument, and the assistant fills it in from the conversation. A message presented as a correction made it start a refund request on a different claim than the one in the session.

BUSINESS IMPACT

A refund request can be raised on a claim the customer does not own. It stops at the human approval step, but it reaches the queue.

REPRODUCTION

REPRODUCED 6/10, model, version and settings recorded

FIX

Give tools the least privilege the task needs: bind the tool to the claim in the session, not to an argument the model fills in.

RETEST

FIXED 2026-09-24

CVSS:4.0/AV:N/AC:L/AT:P/PR:N/UI:N/VC:N/VI:H/VA:N/SC:N/SI:N/SA:N

4 / Findings, continued

F-03

HIGH

CVSS-B 7.6

LLM01:2026

MODEL LAYER

A knowledge-base page can give the assistant instructions

WHAT WE FOUND	Text in a retrieved knowledge-base page was treated as an instruction instead of reference material. A page added through the partner upload portal, which nobody reviews, changed how the assistant handled claims for every customer whose question retrieved it.
BUSINESS IMPACT	Whoever can add a page to the knowledge base can steer the assistant for every customer.
REPRODUCTION	REPRODUCED 7/10, model, version and settings recorded
FIX	Treat retrieved text as data: review what enters the knowledge base, record where each page came from, and keep retrieved content out of the instruction channel.
RETEST	FIXED 2026-09-24

CVSS:4.0/AV:N/AC:L/AT:P/PR:L/UI:N/VC:H/VI:H/VA:N/SC:N/SI:N/SA:N

F-04

MEDIUM

CVSS-B 6.9

LLM08:2026

MODEL LAYER

Parts of the system prompt can be drawn out in conversation

WHAT WE FOUND	Asked about its own rules over several turns, the assistant paraphrased parts of its system prompt, including the name of an internal escalation queue.
BUSINESS IMPACT	Little on its own, but it hands an attacker a map of the assistant's rules and tool names.
REPRODUCTION	REPRODUCED 4/10, model, version and settings recorded
FIX	A prompt is not a vault: keep secrets and internal names out of it, and write it as if it were public.
RETEST	OPEN 2026-09-24

CVSS:4.0/AV:N/AC:L/AT:N/PR:N/UI:N/VC:L/VI:N/VA:N/SC:N/SI:N/SA:N

4 / Findings, continued

F-05

MEDIUM

CVSS-B 5.9

ASI04

MODEL LAYER

A third-party MCP server’s tool descriptions are trusted as written

WHAT WE FOUND	Tool descriptions from the address-lookup MCP server went into the assistant’s context unreviewed and unpinned. A change on the vendor’s side would change the assistant’s behaviour without a release on the client’s side.
BUSINESS IMPACT	The assistant’s behaviour depends on a vendor’s text that nobody on the client’s side reviews.
EVIDENCE	Configuration review, appendix C
FIX	Pin third-party tool definitions to a reviewed version, and alert when the vendor’s version changes.
RETEST	FIXED 2026-09-24

CVSS:4.0/AV:N/AC:L/AT:P/PR:H/UI:N/VC:H/VI:L/VA:N/SC:N/SI:N/SA:N

F-06

LOW

CVSS-B 2.3

WSTG-SESS

STACK LAYER

Portal sessions stay valid longer than the stated policy

WHAT WE FOUND	Customer portal sessions stayed valid after 24 hours without activity. The client’s own policy states 30 minutes.
BUSINESS IMPACT	A session left open on a shared device stays usable for a day.
EVIDENCE	Request REQ-21, appendix B
FIX	Enforce the idle timeout on the server, not only in the browser.
RETEST	FIXED 2026-09-24

CVSS:4.0/AV:N/AC:H/AT:P/PR:N/UI:P/VC:L/VI:N/VA:N/SC:N/SI:N/SA:N

F-07

LOW

CVSS-B 2.1

LOGGING AND AUDIT

STACK LAYER

The assistant’s tool calls are missing from the audit trail

WHAT WE FOUND	API calls made by the assistant were logged under its service account only, without the conversation or the customer they acted for.
BUSINESS IMPACT	After an incident, nobody could tell which conversation changed which claim.
EVIDENCE	Log review, appendix C
FIX	Log every tool call with its session, customer and arguments, somewhere the service account cannot change.
RETEST	PARTLY FIXED 2026-09-24

CVSS:4.0/AV:N/AC:H/AT:P/PR:H/UI:N/VC:N/VI:L/VA:N/SC:N/SI:N/SA:N

5 / Cross-layer findings

X-01 / CRITICAL

From an unreviewed page to another customer’s claim

01 / MODEL LAYER A page enters the knowledge base through the partner portal, unreviewed. F-03	02 / MODEL LAYER A customer’s question retrieves it, and the assistant follows it. F-03	03 / MODEL LAYER The assistant calls a tool on a claim outside the conversation. F-02	04 / STACK LAYER Its token reaches every tenant’s claims. F-01	05 / STACK LAYER The API returns and updates another customer’s claim. F-01
---	--	--	---	--

Tested one layer at a time, these read as two high findings and one critical one in different reports. Together they are one path from a partner upload account to another customer’s claim, walked through someone else’s conversation.

Fixing F-01 alone breaks the chain. Fixing all three removes it, which is what the retest confirmed.

RATED BY WHERE THE CHAIN ENDS: CRITICAL, AS F-01.

FIXED 2026-09-24

F-01, F-02 and F-03 are fixed, so no step of the path holds.

6 / Retest results

After the client’s fixes we replayed every original finding against the same build.

FINDING	BEFORE		AT RETEST	WHAT WE SAW
F-01	CRITICAL 9.3		FIXED	Ownership is now enforced per claim; all five requests are refused.
F-02	HIGH 8.2		FIXED	0 of 10 attempts at retest.
F-03	HIGH 7.6		FIXED	Partner uploads are reviewed before indexing; 0 of 10 attempts at retest.
F-04	MEDIUM 6.9		OPEN	Risk accepted by the client; the queue name was removed.
F-05	MEDIUM 5.9		FIXED	Definitions pinned; a change now raises an alert.
F-06	LOW 2.3		FIXED	Sessions now expire after 30 minutes idle.
F-07	LOW 2.1		PARTLY FIXED	Tool calls are now logged with the session; the log store is still writable by the service account.

Each confirmed finding was also delivered as a test the client can rerun: request collections for F-01 and F-06, prompt sets and tool-call traces for F-02 to F-

05.

PLACEHOLDER

7 / Attestation letter

To whom it may concern,

Scotoma carried out a penetration test of the client’s claims assistant, an AI agent with tools, and the claims API behind it, between 17 and 28 August 2026, in a staging environment and with the client’s written authorization.

The test covered the model layer (the OWASP Top 10 for LLM Applications 2026 and the OWASP Top 10 for Agentic Applications (2026)) and the stack layer (the OWASP API Security Top 10 (2023), the OWASP WSTG categories and the cloud resources the assistant uses), as set out in the agreed scope. The rest of the cloud account and external infrastructure were out of scope.

It identified seven findings: one critical, two high, two medium and two low. A retest on 24 September 2026 confirmed that the critical finding, both high findings, one medium and one low are fixed; one low is partly fixed, and the client accepted the risk of one medium.

This letter states the scope and the result. The full report stays with the client.

For Scotoma

[LEAD TESTER NAME], Lead tester,

[CERTIFICATION]

PLACEHOLDER