WOOLLEMIA
INFOSEC

SAMPLE REPORT

Agentic AI Security Assessment

A practical security assessment for AI agents, LLM applications, RAG pipelines and MCP-enabled tool ecosystems.

Sample organization	NovaFin AI Platform
Assessment type	Grey-box Agentic AI Security Assessment
Report status	Illustrative sample - not a live assessment
Prepared by	Woollemia Infosec Pvt Ltd

Confidentiality notice: This sample contains fictional systems and findings created to demonstrate the depth, structure and clarity of Woollemia Infosec reporting.

01 / EXECUTIVE VIEW

Executive Summary

The sample environment is an AI support agent that retrieves internal documents and invokes ticketing, customer-record and email tools through an MCP gateway.

Overall risk	5 findings identified	Primary exposure
HIGH	1 Critical | 2 High | 1 Medium | 1 Low	Untrusted content can influence tool-enabled agent actions.

Business impact

A successful attack could cause the agent to disclose sensitive customer data, create unauthorized tickets, send externally controlled messages or retain malicious instructions in long-term memory. The most important control gap is that retrieved content and tool output are treated as trusted instructions instead of untrusted data.

Priority actions

Priority	Action	Expected outcome
P0	Require explicit authorization and human approval for high-impact tool actions.	Stops silent privilege use and destructive actions.
P0	Separate trusted system instructions from retrieved content and tool output.	Reduces indirect prompt injection success.
P1	Enforce record-level authorization outside the model.	Prevents cross-user and cross-tenant data exposure.
P1	Add AI-specific telemetry, abuse cases and response playbooks.	Improves detection and containment.

Scope, Architecture & Method

Testing combines architecture review, manual adversarial testing and safe tool-action validation. Production data is not modified.

In scope	Coverage
AI agent	System prompt, planning, memory, output processing and refusal behavior
RAG pipeline	Document ingestion, retrieval authorization, source trust and cross-tenant isolation
MCP gateway	Server trust, tool descriptions, credential boundaries and action authorization
Connected tools	CRM lookup, ticket creation and outbound email - safe non-destructive validation
AI API layer	Authentication, rate controls, logging, secrets and error handling

Assessment lenses

Model behavior: direct and indirect prompt injection, prompt leakage, unsafe output handling and denial-of-service patterns.
Agent behavior: goal hijacking, excessive agency, memory poisoning, unsafe planning and approval bypass.
Application controls: identity, authorization, tenant separation, secrets, telemetry, rate controls and incident response.
Ecosystem trust: RAG source integrity, MCP server trust, tool definitions, dependency exposure and supply-chain risk.

Reference frameworks

OWASP Top 10 for LLM Applications, OWASP Agentic AI guidance, MITRE ATLAS, NIST AI RMF and ISO/IEC 42001 control themes.

Findings Overview

Severity reflects exploitability, required access, affected data or actions, blast radius and the reliability of existing safeguards.

ID	Finding	Severity	Status
AI-01	Indirect prompt injection triggers unauthorized tool actions	Critical	Open
AI-02	Cross-tenant retrieval authorization bypass	High	Open
AI-03	Excessive MCP tool permissions	High	Open
AI-04	Persistent memory poisoning	Medium	Open
AI-05	Verbose model and connector errors	Low	Open

Attack path observed

1 Attacker places a malicious instruction inside a document.	**2** RAG retrieves the document as context for the agent.	**3** Agent interprets content as an instruction and selects a tool.	**4** MCP gateway executes the action without sufficient approval.

CRITICAL

AI-01: Indirect Prompt Injection Triggers Unauthorized Tool Actions

Affected components: RAG ingestion, AI agent planner, MCP email tool

Description

The agent followed instructions embedded in an untrusted knowledge-base document. When the document was retrieved to answer a user query, the model treated the embedded text as a higher-priority operational instruction and invoked an outbound-email tool. The action was accepted without a clear user confirmation step.

Illustrative proof of concept

Test input	Knowledge document containing hidden instructions to summarize customer data and send it to a supplied address.
Observed behavior	The agent retrieved the document, changed its plan and prepared an external tool call using data from the active conversation.
Control failure	Retrieved content was not marked as untrusted; the tool gateway did not require action-specific approval.

Impact

An attacker may cause unauthorized disclosure, impersonation, fraudulent workflow execution or destructive downstream actions. The impact increases when the agent holds broad credentials or acts across multiple tenants.

Recommendations

1. Treat all retrieved and tool-returned content as untrusted data, never as executable instructions.
2. Require explicit user confirmation for external communications, financial actions, record changes and other high-impact tools.
3. Bind tool authorization to the authenticated user, requested purpose and target resource.
4. Apply allowlisted tool parameters and deterministic policy checks outside the LLM.
5. Log the full decision chain and alert on instruction-source conflicts.

HIGH

AI-02: Cross-Tenant Retrieval Authorization Bypass

A user from Tenant A could influence retrieval with document identifiers associated with Tenant B. The retrieval service trusted model-generated filters and did not independently bind the query to the caller tenant.

Risk

Sensitive contracts, support records or internal knowledge may be disclosed across tenants through ordinary natural-language queries.

Key remediation

Enforce tenant and record authorization in the retrieval layer using server-side identity context. Do not allow the model to create or override security filters. Add negative cross-tenant regression tests.

HIGH

AI-03: Excessive MCP Tool Permissions

The MCP connector credential permitted read, create, update and delete operations although the business workflow only required limited read and create access.

Risk

Any successful prompt injection, agent planning error or compromised MCP server receives a larger operational blast radius.

Key remediation

Issue per-tool, per-environment credentials; remove unused operations; restrict resource scope; add short-lived tokens; and require policy checks before sensitive actions.

Supporting Findings & Control Mapping

ID	Observation	Severity	Recommended control
AI-04	A malicious instruction could persist in agent memory and influence later sessions.	Medium	Provenance, expiry, sanitization and review for memory writes.
AI-05	Verbose errors exposed model, connector and internal object details.	Low	Generic user errors; secure server-side diagnostic logging.

Illustrative framework mapping

Finding	OWASP LLM risk theme	MITRE ATLAS theme	NIST AI RMF outcome
AI-01	Prompt injection / excessive agency	LLM prompt injection	Manage AI risks and monitor behavior
AI-02	Sensitive information disclosure	Data from information repositories	Govern access and protect data
AI-03	Excessive agency / supply chain	Abuse external services	Map dependencies and manage controls
AI-04	Data and model poisoning	Poison training or operational data	Measure integrity and monitor drift

Prioritized Remediation Roadmap

The roadmap sequences immediate containment, engineering fixes and longer-term assurance so risk can be reduced without blocking responsible AI adoption.

Timeline	Focus	Actions
0-7 days	Contain critical paths	Disable unnecessary tools; add approvals for high-impact actions; isolate vulnerable knowledge sources; rotate over-privileged credentials.
2-4 weeks	Strengthen architecture	Enforce retrieval authorization; introduce policy-as-code for tools; separate instructions from data; implement memory provenance.
30-60 days	Operationalize assurance	Add AI-specific logging, alerts, abuse-case tests, red-team regression suites and incident-response playbooks.
Quarterly	Maintain resilience	Re-assess high-risk workflows after model, tool, MCP server or architecture changes.

Re-assessment and closure

Woollemia Infosec validates implemented fixes using the original evidence path and targeted regression tests. The final report records the remediation state, residual risk, retest evidence and closure status for each finding.

Client-ready deliverables

Executive summary
Detailed findings with evidence
Architecture observations
Prioritized remediation roadmap
Re-assessment results
Final letter of attestation, where applicable

Why Woollemia Infosec

Manual plus tool-assisted testing
Safe business-logic validation
Engineering-friendly remediation
Revalidation support
Auditor-ready reporting
Long-term security partnership

WOOLLEMIA
INFOSEC

Build powerful AI. Keep every action accountable.

Woollemia Infosec helps organizations test the security of AI agents, LLM applications, RAG systems, MCP integrations and the APIs and cloud controls around them.

Request an assessment
info@woollemia.com
www.woollemia.com

This document is a fictional sample. Actual findings, risk ratings, evidence and recommendations are tailored to the assessed architecture, agreed scope and observed controls.