CASE STUDY

Post-Compromise Impact Assessment for Newly Built Platform Agents

2025

sidekicksecurity.io

EXECUTIVE SUMMARY

> *Across eight distinct attack paths, spanning prompt injection, authentication lifecycle management, permission scoping, session persistence, search access verification, custom agent governance, and runtime behavioral controls, Sidekick Security demonstrated that even well-architected AI agents in mature enterprise-ready platforms carry compounding vulnerabilities that only a threat-model-driven assessment methodology can reveal.*

Organizations deploying AI agents face a uniquely challenging security landscape, whether as standalone functioning agents or as part of a broader SaaS platform. The convergence of natural language interfaces, LLM-powered reasoning, dynamic tool execution, and enterprise data access creates attack surfaces that traditional security assessments are not designed to evaluate. In this case study, we examine how applying a post-compromise threat modeling methodology, purpose-built for AI agent architectures, reveals cascading attack paths that no single-domain assessment can detect.

The body of work reveals a consistent theme: organizations that invest in foundational AI security controls such as schema validation and permission enforcement remain exposed when those controls are not stress-tested under realistic adversarial conditions. When authentication lifecycle gaps, over-provisioned credentials, and unmonitored session persistence exist alongside prompt injection exposure, each weakness compounds the others; creating attack chains that move from a single malicious input source to full data exfiltration within a single exploitation window.

As highlighted throughout this document, Sidekick's methodology bridges this gap by analyzing every layer an attacker would target, chaining findings into compound attack scenarios, and delivering deployable remediations that close gaps immediately rather than leaving teams with a report and no clear path forward.

1	4	8	5	3
ORGANIZATION	AI AGENTS IN SCOPE	ATTACK PATHS IDENTIFIED	CRITICAL BUSINESS RISKS	PHASED RESPONSE PLANS

CLIENT PROFILE

Industry	Enterprise Technology: AI-Powered Data Platform
Organization Scope	Cloud-native AI agent platform integrated into enterprise data platform, covering agent orchestration infrastructure, LLM gateway integrations, session management, and custom agent creation capabilities
Geographic Footprint	AI agent architecture (v1 and v2), authentication lifecycle, permission inheritance and enforcement, session persistence, custom agent governance, and runtime behavioral controls
Regulatory Environment	ISO 27001, NIST Cybersecurity Framework (CSF), SOC 2, FedRAMP, and third-party LLM data handling obligations
Assessment In-Scope Domains	AI agent code repositories and design documentation, data access control mechanisms, trust boundaries between agents and platform services, integration points, separation of duties controls, and user query processing flows

THE CHALLENGE

Enterprises building AI agent capabilities are managing security risk in a quickly changing landscape where established frameworks and tooling have not yet caught up with the threat surface. AI agents, particularly those undergoing architectural evolution from simple single-call patterns to sophisticated multi-step agentic workflows, introduce attack vectors that traditional application security assessments are not designed to evaluate.

Modern AI agent platforms typically involve an orchestration layer managing dynamic tool execution, authentication credentials and identity context shared across service calls, natural language inputs processed as trusted instructions, and persistent session data that outlives individual interactions. Each represents a distinct attack surface that adversaries can chain together. The need for a dedicated, adversarial threat modeling methodology, focused on post-compromise impact and blast radius, drove the engagement documented in this case study.

Key Concerns

- AI agent architecture undergoing significant evolution from single-call v1 agents to multi-step v2 agentic workflows, expanding the attack surface faster than security controls could be validated
- Prompt injection detection planned but not yet deployed, leaving agents exposed to manipulation through malicious content, shared resource poisoning, and direct input attacks across all agent modalities
- Internal JWT tokens valid for 60 minutes with no revocation mechanism, making it impossible to immediately terminate a compromised session even when an active attack is detected by the security team

- JWT downscoping not implemented: every agent tool call receives the invoking user's full account credentials regardless of the scope or intent of the task being performed
- AI sessions persisting indefinitely without mandatory expiration policies, enabling data captured during an attack to remain accessible long after the active exploitation window closes
- Custom agent creation available to all platform users without approval workflows, instruction validation, or content scanning; enabling malicious insider or compromised account scenarios
- No guardrail APIs deployed to provide runtime behavioral controls, leaving agents without a standard defensive layer designed specifically for LLM-based attack techniques
- Absence of behavioral monitoring baselines for agent-to-service call patterns, limiting the ability to detect active attacks or post-compromise lateral movement across internal services

THE APPROACH

Sidekick Security conducted a three-week AI Agent Architecture Risk Analysis focused on threat modeling and post-compromise impact assessment. The engagement was scoped to the organization's AI agent ecosystem, covering multiple agents sharing a similar development stack, and was designed around an assumed-breach scenario: evaluating the cascading effects and blast radius of a successful prompt injection or other attack against the agent platform.

The approach combined architecture discovery and source code analysis with deep threat modeling and blast radius quantification, ensuring that both exploitable vulnerabilities and structural design weaknesses were identified and contextualized within the product engineering environment. Every finding was mapped to applicable compliance frameworks, and all remediations were designed for direct integration into product roadmaps.

Activities Performed

SERVICE	DESCRIPTION
Architecture Discovery & Attack Surface Mapping	Review of agent architecture documentation, technology stack analysis, examination of integration patterns with platform APIs, and mapping of trust boundaries, permission contexts, and data flow patterns across the v1 and v2 agent ecosystem
Deep Threat Modeling (Post-Compromise Focus)	Identification of realistic compromise vectors, analysis of permission boundary violations and privilege escalation paths, assessment of lateral movement and cross-service impact scenarios, data exfiltration modeling, and evaluation of persistence mechanisms
Architectural Security Control Validation	Targeted source code analysis to validate identified issues, review of permission enforcement implementation, analysis of logging and monitoring capabilities, and assessment of containment controls that limit blast radius

Blast Radius & Impact Quantification	For each identified attack chain: quantification of immediate access gained, privilege level achieved, user impact scope, volume of data exposed, potential for lateral movement, likelihood of detection, containment effectiveness, and business impact severity
Compliance Impact Analysis	Mapping of each identified vulnerability to applicable regulatory frameworks (ISO 27001, NIST CSF, FedRAMP, SOC 2) with gap analysis, compliance consequence assessment, and audit-ready remediation documentation
Detection & Response Strategy Development	Three-phased enhancement roadmap covering immediate detection logging, enhanced behavioral monitoring, and automated response automation, with deployable detection logic for each identified attack phase

KEY FINDINGS

Across the AI agent architecture assessment, Sidekick Security identified eight attack paths spanning prompt injection exposure, authentication lifecycle weaknesses, permission model gaps, session persistence risks, and governance deficiencies in custom agent creation. The findings below represent the most impactful vulnerabilities identified, each carrying significant blast radius when combined with the interconnected weaknesses present across the architecture.

Prompt Injection & Runtime Controls

HIGH

Prompt Injection Detection Not Deployed

Prompt injection detection was planned but not yet deployed at the time of assessment. Agents process all content, various user input sources, metadata, file names, hidden text, and shared resource content as trusted input without scanning for injection payloads. This absence is the root cause vulnerability enabling most other attack scenarios: without the ability to manipulate agent behavior, attackers cannot exploit JWT persistence gaps, over-provisioned permissions, or session artifacts. Every other finding in this assessment is amplified by this gap.

HIGH

Absence of Guardrail APIs

No guardrail API, such as LLaMA Guard or Nemo Guardrails, was deployed to provide runtime behavioral controls for agent execution. This removes an entire defensive layer considered standard practice in LLM application security. Without guardrails, the platform has no runtime mechanism to detect prompt injection attempts, jailbreak patterns, requests for harmful content generation, or anomalous tool invocation sequences. Certain attack categories face reduced resistance across every agent interaction.

Authentication & Session Management

HIGH

No Internal JWT Token Revocation Capability

Internal JWTs used for all agent tool calls are valid for 60 minutes with no mechanism to revoke them before expiration. When a security team detects an active compromise, they cannot terminate the session; they must either wait up to 60 minutes for natural expiration or resort to disruptive measures such as disabling the victim's user account entirely. This gap directly limits incident response effectiveness and extends the exploitation window during active attacks, regardless of how quickly the compromise is detected.

HIGH

JWT Downscoping Not Implemented

All agent tool calls, regardless of task scope, receive the invoking user's full credentials, granting access to the user's entire account rather than only the content relevant to the specific task at hand. When a user asks an agent to summarize a selection of input text, the agent receives credentials capable of accessing every resource the user can reach. This over-provisioning directly amplifies the blast radius of any successful prompt injection or agent compromise, turning a narrow task into a broad data exposure opportunity.

MODERATE

Session Persistence and Shared Session Risks

AI sessions persist indefinitely without mandatory expiration policies. Session data, including search results, data summaries, and execution history, remains accessible until manually deleted or until an administrator policy forces expiration. Sharing a session grants recipients access to generated content derived from files they cannot directly access, creating an authorization bypass. Compromised sessions serve as long-term data retention artifacts accessible well after the active attack window closes.

Data Access & Agent Governance

CRITICAL

Prompt Injection to Internal Service Lateral Movement

A successful prompt injection attack can cascade into lateral movement across the platform's internal service architecture. Because all agent tool calls trust the user's JWT without additional per-service validation, a compromised agent gains indirect access to the intelligence layer, LLM gateway, search APIs, and metadata services, all operating under the victim's credentials. With prompt injection detection absent and no behavioral monitoring baseline established, the likelihood of detection during active exploitation is critically low.

HIGH

Insufficient Search Result Permission Verification

Agent search operations may retrieve and process user-provided content without re-verifying user permissions at retrieval time. A compromised agent can manipulate search filters, removing date ranges, expanding scope, or targeting specific content types, to access content beyond the user's intended scope. Even where the underlying search API limits results to content the user has technical access to, manipulated searches expose content the user never intended to provide to the agent for the current task.

HIGH

Custom Agent Creation and Configuration Risks

Custom agents can be created by any authenticated user, not just administrators, without approval workflows, custom instruction validation, or content scanning for malicious payloads. An attacker can create agents with embedded instructions that affect all users who interact with them, attach data sources containing prompt injection payloads, or use agents as data exfiltration mechanisms. An identified gap in sharing controls allows agents to be distributed externally without triggering standard administrative oversight.

KEY THEMES & TRENDS

Applying adversarial threat modeling to an AI agent architecture surfaces patterns across attack paths, components, and risk categories that narrow-scope assessments cannot reveal. The following themes emerged consistently across this engagement, each representing a systemic challenge that organizations deploying AI agents must address through coordinated security programs rather than isolated, point-solution remediation efforts.

8/8	60 MIN	3
ATTACK PATHS CHAIN ACROSS COMPONENTS	UNCONTROLLABLE JWT EXPOSURE WINDOW	EFFECTIVE CONTROLS IDENTIFIED

1

Prompt Injection Is the Root Cause That Enables Every Other Attack

The assessment consistently confirmed that prompt injection is not merely one vulnerability among many, it is the root cause that unlocks every other attack path. Without prompt injection detection, all other architectural weaknesses become immediately exploitable. Organizations that treat AI-specific input validation as a secondary concern after platform functionality is complete are taking on compounding risk that grows with every additional agent capability deployed.

2

Authentication Lifecycle Gaps Create Uncontrollable Incident Response Windows

The inability to revoke internal JWT tokens before their 60-minute expiration represents one of the most operationally impactful findings in this assessment. Security teams facing an active compromise cannot terminate the session, the only immediate options are waiting out the attack window or disabling the victim's account. This gap inverts normal incident response, forcing organizations to choose between prolonged exposure and business disruption.

3

Over-Provisioned Permissions Transform Narrow Tasks Into Broad Exposures

Agent platforms that issue full-scope credentials for every tool call, regardless of task context, systematically amplify the blast radius of any successful attack. When an agent asked to summarize a user input receives credentials for an entire account, any compromise of that agent is effectively a full-account compromise. This pattern was identified as a consistent architectural gap across the AI agent implementations reviewed.

4

Custom Agent Governance Lags Behind Platform Capability Expansion

As platforms expand agent creation capabilities from administrators to all users, corresponding governance controls do not scale proportionally. The result is a growing population of user-created agents with unvalidated custom instructions, unscanned data sources, and inadequate sharing controls, each representing a potential attack vector for malicious insiders or compromised accounts. The gap between what users can create and what the security team can monitor widens with every platform release.

5

Session Persistence Creates Long-Tail Data Exposure That Outlasts Active Attacks

Session data that persists indefinitely extends the effective blast radius of any successful compromise far beyond the active JWT window. An attacker who extracts data during a 60-minute session retains permanent access to those session artifacts. Combined with the authorization bypass created by sharing generated content derived from restricted files, session persistence transforms a time-bounded attack into a long-term data exposure event.

6

AI-Specific Defense-in-Depth Requires Controls That Traditional Security Tools Cannot Provide

Organizations that rely on perimeter controls, schema validation, and platform-layer permission enforcement as their primary AI security posture significantly overestimate their resilience against AI-native attack techniques. Guardrail APIs, behavioral baseline monitoring, prompt injection detection, and LLM-specific output filtering are not incremental improvements to existing controls, they are the foundational layer that makes AI agent security measurable and defensible.

CROSS-DOMAIN VULNERABILITY CHAINING

The most significant insight from an AI agent threat model assessment is how vulnerabilities across independent components combine to create compound attack scenarios with exponentially greater impact. Sidekick's methodology deliberately maps these cross-component chains, the same approach real adversaries use, revealing risks that single-vector analysis fundamentally cannot detect.

Full-Spectrum Attack Chain

STEP 1
Injection Entry
Malicious payload delivered via indirect or direct input

▸

STEP 2
Agent Hijack
Prompt injection overrides agent behavior & intent

▸

STEP 3
JWT Exploitation
Full-scope credentials used for lateral service access

▸

STEP 4
Data Exfiltration
Enterprise content accessed, extracted & persisted

Chain 1: Prompt Injection to Internal Service Lateral Movement

DELIVERY	An attacker crafts a prompt injection payload embedded in a source uploaded to the platform, using hidden text, metadata fields, file names, or shared resource content. The payload instructs the agent to perform actions beyond the user's actual request. With no content scanning at ingestion and no prompt injection detection deployed, the payload reaches the agent unimpeded.
INGESTION	The victim user invokes an AI agent to perform a legitimate task, such as analyzing or summarizing the input source. The agent processes all content as trusted instruction. The embedded payload redirects the agent's behavior without the user's awareness, causing it to begin executing the attacker's instructions rather than fulfilling the user's request.
ESCALATION	The compromised agent begins executing attacker-directed operations using the victim's full JWT. Because downscoping is not implemented, those credentials grant access to every resource, service, and data store the user can reach, far beyond the single resource originally requested. The agent traverses the intelligence service, search APIs, and metadata services under the victim's identity.
LATERAL MOVEMENT	The agent pivots across internal services: issuing broad searches, accessing metadata, querying content categories the user never requested. Each service trusts the user's JWT without additional validation. No behavioral monitoring exists to flag the anomalous cross-service access pattern as it unfolds in real time.
PERSISTENCE	Even after the active exploitation window closes, session data, containing search results, summaries, and execution history, persists indefinitely. Because JWTs cannot be revoked, the security team cannot terminate the session during the attack. The compromise extends beyond the 60-minute window through session artifacts that remain accessible until explicitly deleted.

Chain 2: JWT Exploitation for Persistent Post-Detection Access

INITIAL ACCESS	The attacker gains control of an agent session through prompt injection, malicious content ingestion, or session manipulation. Upon compromise, the attacker's commands execute using the victim's JWT automatically, the compromised agent inherits all credentials without requiring the attacker to extract tokens directly.
CONTINUED ACCESS	Using the compromised session, the attacker has up to 60 minutes of access to all internal services accepting the user's JWT. Critically, this window persists even if the user logs out of their external web session, because internal JWTs operate independently from the user's primary authentication context.
DETECTION EVASION	If the security team detects the attack and attempts to terminate the session, they currently have no mechanism to revoke internal JWTs. External session IDs can be invalidated, but the underlying JWT continues to authorize internal service calls. The

	attacker's access persists until the token expires, regardless of what response actions the security team takes.
IMPACT	For users with elevated permissions, administrators, executives, or users with broad content access, the 60-minute window provides a substantial opportunity to enumerate sensitive content, extract organizational data, and establish artifacts that persist beyond the active window. Disabling the user's account is the only effective containment, and carries its own operational disruption.

Chain 3: Custom Agent Poisoning to Organization-Wide Propagation

CREATION	An attacker, either a malicious insider or an external actor who has compromised an account, creates a custom agent through the platform's agent creation interface. The agent is given a legitimate name and description to encourage adoption. Custom instructions embedding malicious behaviors are not validated or scanned before the agent is activated.
POISONING	The attacker attaches data sources containing prompt injection payloads. When users query the agent, the poisoned content is retrieved and processed, triggering the payload under that user's permissions. Unlike direct prompt injection requiring the attacker to target each victim individually, poisoned data sources allow a single malicious file to affect every user who interacts with the agent.
DISTRIBUTION	The agent is shared broadly, within the organization or externally. Because the agent management folder exists outside of global administrative sharing controls, external distribution does not trigger standard oversight mechanisms. Recipients interact with what appears to be a helpful, legitimate productivity tool with no visible indication of malicious configuration.
EXFILTRATION	Each victim's interaction executes under their unique permissions. Malicious custom instructions cause the agent to collect sensitive content, access organizational data, and include extracted information in its responses. Shared sessions grant access to generated content derived from files recipients cannot directly view, effectively laundering sensitive data into shareable form.
IMPACT	If the malicious agent is distributed broadly, hundreds of users could be affected, with each interaction compounding the data exposure. Because custom instructions are not visible to end users and no behavioral monitoring baseline exists for custom agent execution, the campaign could operate for an extended period before detection through manual audit or downstream incident.

WHY FULL-SPECTRUM CHAINING MATTERS

Each vulnerability in these chains might receive a "high" or "moderate" rating in isolation. But when chained across the agent architecture, they create realistic paths from a single malicious data source to full organizational data exposure, within a 60-minute JWT window that cannot be revoked, with session artifacts that persist indefinitely. Sidekick's methodology ensures these compound risks are identified, and documented as interconnected attack paths, not as disconnected findings scattered across separate reports.

This approach also highlights the critical importance of monitoring and response capabilities that operate across the entire agent execution lifecycle.

DEPLOYABLE REMEDIATIONS

Sidekick Security goes beyond identifying vulnerabilities; every engagement produces ready-to-deploy detection rules, architectural guidance, governance frameworks, compliance mappings, and phased response roadmaps. These deliverables close the gap between "findings" and "fixed" immediately, rather than leaving organizations with a report and no actionable path forward.

Prompt Injection Detection

Deploy prompt injection detection as a critical near-term priority. Complement with content scanning at ingestion, inspecting uploads for known payload patterns, hidden text techniques (white-on-white, small fonts), metadata field injection, and anomalous content structures before agents process any content.

Guardrail API Deployment

Deploy a guardrail API solution, LLaMA Guard, Nemo Guardrails, or equivalent, to provide runtime behavioral protection. These systems detect and block prompt injection attempts, jailbreak patterns, and anomalous output behaviors as a pre/post-processing layer integrable without significant architectural changes.

JWT Revocation & Downscoping

Implement emergency JWT revocation capability; the most critical gap for incident response. Additionally, implement JWT downscoping per tool call, ensuring agents receive credentials scoped only to the specific resource and task at hand. Even if an agent is compromised, attackers can only access resources within the narrow token scope.

Session Lifecycle Management

Implement mandatory session expiration policies with sensitivity-based maximum lifetimes (high sensitivity: 1 hour; medium: 4 hours; low: 24 hours). Add content sensitivity scanning for session data, force expiration on anomaly detection, and alert on unusually long or high-volume sessions.

Custom Agent Governance Framework

Implement a validation and approval framework for custom agent configurations. Scan custom instructions for malicious patterns. Require administrator approval before broad sharing. Classify data sources for sensitivity before attachment. Bring agent sharing under standard administrative controls to eliminate the external sharing gap.

Behavioral Monitoring & Baselines

Establish behavioral baselines for normal agent operations; search volumes, file access patterns, tool invocation sequences, and cross-service call correlations. Alert on deviations. Implement SIEM-based cross-service correlation across the intelligence layer, search APIs, and LLM gateway to identify lateral movement patterns indicative of a compromised session.

LLM Provider Contractual Controls

Maintain explicit contractual agreements with all LLM providers prohibiting use of customer data for model training. Specify that prompts, content, and extracted information must remain confidential and cannot be used to train or improve any models.

Search Permission Re-Verification

Implement permission re-verification for search results at retrieval time, ensuring that even if search filters are manipulated, agents can only process inputs appropriate for the current task context. Add task-context-aware search scope controls that constrain queries based on the user's stated intent.

Detection Rule Examples

The following tables illustrate the type of deployable detection logic Sidekick delivers as part of every engagement. Rules are tailored to the client's specific SIEM platform, log sources, and operational context.

Agent Behavior — Anomalous Reconnaissance and Search Abuse

Rule Name	AI-001: Anomalous Agent Search Abuse — Reconnaissance Pattern
Trigger	Single agent session initiates searches for more than 100 results within a session; OR search filters are removed from baseline query patterns; OR agent queries span more than 3 unrelated content categories within 5 minutes of session start Application-side search logging: {timestamp, user_id, session_id, query_text, filters_applied, result_count, data_accessed} High: escalate to Critical if search targets content categories inconsistent with the session-initiating query (e.g., a summarization task pivoting to broad organizational searches) Alert security team, flag session for review, correlate with recent upload activity to identify potential injection source, preserve session artifacts for forensic analysis
Data Sources	Single agent session initiates searches for more than 100 results within a session; OR search filters are removed from baseline query patterns; OR agent queries span more than 3 unrelated content categories within 5 minutes of session start Application-side search logging: {timestamp, user_id, session_id, query_text, filters_applied, result_count, data_accessed} High: escalate to Critical if search targets content categories inconsistent with the session-initiating query (e.g., a summarization task pivoting to broad organizational searches) Alert security team, flag session for review, correlate with recent upload activity to identify potential injection source, preserve session artifacts for forensic analysis
Severity	Single agent session initiates searches for more than 100 results within a session; OR search filters are removed from baseline query patterns; OR agent queries span more than 3 unrelated content categories within 5 minutes of session start Application-side search logging: {timestamp, user_id, session_id, query_text, filters_applied, result_count, data_accessed} High: escalate to Critical if search targets content categories inconsistent with the session-initiating query (e.g., a summarization task pivoting to broad organizational searches) Alert security team, flag session for review, correlate with recent upload activity to identify potential injection source, preserve session artifacts for forensic analysis
Response	Single agent session initiates searches for more than 100 results within a session; OR search filters are removed from baseline query patterns; OR agent queries span more than 3 unrelated content categories within 5 minutes of session start Application-side search logging: {timestamp, user_id, session_id, query_text, filters_applied, result_count, data_accessed} High: escalate to Critical if search targets content categories inconsistent with the session-initiating query (e.g., a summarization task pivoting to broad organizational searches) Alert security team, flag session for review, correlate with recent upload activity to identify potential injection source, preserve session artifacts for forensic analysis

Cloud — Azure Security Group Manipulation

Rule Name	AZ-001: Unauthorized Security Group Creation
Trigger	Non-admin user creates a new security group in Azure AD; OR security group membership changes by non-privileged account; OR new group assigned to Key Vault or storage account access policy
Data Sources	Azure AD Audit Logs, Azure Activity Logs, Microsoft Sentinel
Severity	Moderate (escalate to High if group gains access to Key Vault or sensitive resources)
Response	Alert security team, review group membership and assigned permissions, verify business justification

Authentication — JWT Session Anomaly Detection

Rule Name	AI-002: Lateral Movement — Agent Cross-Service Access Pattern
Trigger	Single agent session accesses more than 3 distinct internal services within 5 minutes; OR internal service calls continue after the user's external session has been terminated; OR agent session duration exceeds 8 hours without explicit user activity Application-side service call logging: {timestamp, user_id, session_id, service_name, action_type, resource_ids, token_scope} Moderate: escalate to High if cross-service access pattern includes the LLM gateway alongside broad search activity, or if session persists post-logout Alert security team, review session artifact contents, initiate JWT revocation if capability is available, assess whether sensitive content was captured in session history
Data Sources	Single agent session accesses more than 3 distinct internal services within 5 minutes; OR internal service calls continue after the user's external session has been terminated; OR agent session duration exceeds 8 hours without explicit user activity Application-side service call logging: {timestamp, user_id, session_id, service_name, action_type, resource_ids, token_scope} Moderate: escalate to High if cross-service access pattern includes the LLM gateway alongside broad search activity, or if session persists post-logout Alert security team, review session artifact contents, initiate JWT revocation if capability is available, assess whether sensitive content was captured in session history
Severity	Single agent session accesses more than 3 distinct internal services within 5 minutes; OR internal service calls continue after the user's external session has been terminated; OR agent session duration exceeds 8 hours without explicit user activity Application-side service call logging: {timestamp, user_id, session_id, service_name, action_type, resource_ids, token_scope} Moderate: escalate to High if cross-service access pattern includes the LLM gateway alongside broad search activity, or if session persists post-logout Alert security team, review session artifact contents, initiate JWT revocation if capability is available, assess whether sensitive content was captured in session history
Response	Single agent session accesses more than 3 distinct internal services within 5 minutes; OR internal service calls continue after the user's external session has

	been terminated; OR agent session duration exceeds 8 hours without explicit user activity Application-side service call logging: {timestamp, user_id, session_id, service_name, action_type, resource_ids, token_scope} Moderate: escalate to High if cross-service access pattern includes the LLM gateway alongside broad search activity, or if session persists post-logout Alert security team, review session artifact contents, initiate JWT revocation if capability is available, assess whether sensitive content was captured in session history

Custom Agent — High-Risk Creation and Propagation

Rule Name	AI-003: Custom Agent Propagation — Suspicious Creation and Sharing
Trigger	Non-admin user creates an agent with custom instructions exceeding 500 characters; OR newly created agent is shared with more than 10 users within 24 hours; OR agent is shared externally within any time window; OR agent data sources include content classified as sensitive Agent configuration logging with alerting: creation events, collaboration/sharing audit logs, content classification service events Moderate: escalate to High if agent is shared externally, or if instruction content matches known injection signatures or data exfiltration command patterns Route to administrator review queue, hold agent activation pending approval, notify security team of external sharing attempt, scan instruction content against known malicious patterns
Data Sources	Non-admin user creates an agent with custom instructions exceeding 500 characters; OR newly created agent is shared with more than 10 users within 24 hours; OR agent is shared externally within any time window; OR agent data sources include content classified as sensitive Agent configuration logging with alerting: creation events, collaboration/sharing audit logs, content classification service events Moderate: escalate to High if agent is shared externally, or if instruction content matches known injection signatures or data exfiltration command patterns Route to administrator review queue, hold agent activation pending approval, notify security team of external sharing attempt, scan instruction content against known malicious patterns
Severity	Non-admin user creates an agent with custom instructions exceeding 500 characters; OR newly created agent is shared with more than 10 users within 24 hours; OR agent is shared externally within any time window; OR agent data sources include content classified as sensitive Agent configuration logging with alerting: creation events, collaboration/sharing audit logs, content classification service events Moderate: escalate to High if agent is shared externally, or if instruction content matches known injection signatures or data exfiltration command patterns Route to administrator review queue, hold agent activation pending approval, notify security team of external sharing attempt, scan instruction content against known malicious patterns
Response	Non-admin user creates an agent with custom instructions exceeding 500 characters; OR newly created agent is shared with more than 10 users within 24 hours; OR agent is shared externally within any time window; OR agent data sources include content classified as sensitive Agent configuration logging with alerting: creation events, collaboration/sharing audit logs, content classification service events

	Moderate: escalate to High if agent is shared externally, or if instruction content matches known injection signatures or data exfiltration command patterns Route to administrator review queue, hold agent activation pending approval, notify security team of external sharing attempt, scan instruction content against known malicious patterns

GRC Impact Summary

Every engagement produces comprehensive compliance documentation, mapping findings to applicable regulatory frameworks and risk register entries. This helps our customers not only close real gaps, but stay prepared for regulatory and external audits that put pressure on the team. The table below demonstrates how Sidekick contextualizes findings from all assessment domains within the regulatory environment specific to credit unions and financial institutions.

FRAMEWORK	CONTROL	FINDING IMPACT	REMEDIATION
ISO 27001	A.9.4.1, A.12.4.1, A.13.1.1	Prompt injection enables unauthorized access to internal services, violating access control (A.9.4), monitoring (A.12.4), and network security (A.13.1) requirements	Deploy prompt injection detection, guardrail APIs, and behavioral monitoring with cross-service correlation
NIST CSF	PR, RS	Inability to revoke active sessions extends attack windows up to 60 minutes, violating access termination (PR.AC) and incident containment (RS.MI) requirements.	Implement JWT revocation capability, reduce token lifetime, deploy real-time session monitoring and anomaly alerting
NIST CSF	PR	Over-provisioned agent credentials violate least privilege (PR.AC-4) and data protection (PR.DS-5) principles, enabling access well beyond task scope	Implement JWT downscoping per tool call with task-context-aware scope boundaries
SOC2	CC6, CC7	Absence of input validation and malicious content detection creates gaps in continuous monitoring (CC7.1) and system operations (CC7.2) controls	Deploy guardrail APIs (LLaMA Guard or Nemo Guardrails) with content scanning at input ingestion
SOC 2	CC6	Indefinite session persistence and derived content sharing create authorization bypass vectors, violating logical access (CC6.1) and data transmission controls (CC6.7)	Implement mandatory session expiration with sensitivity-based policies and content scanning for session data
ISO 27001	A.9, A.12, A.14	Unrestricted agent creation without validation violates access provisioning (A.9.2), change management (A.12.5), and secure development (A.14.2) requirements	Implement tiered approval workflows, instruction scanning, and administrative controls over agent sharing

VSEC: CONTINUOUS SECURITY PARTNERSHIP

> *A threat model assessment report identifies the gaps. vSec closes them. Sidekick's vSec subscription service embeds our engineers and security leaders directly with your team to accelerate remediation, build program maturity, and transform findings into measurable security improvements… not just a report on a shelf.*

Most organizations complete a security assessment, receive a report full of critical findings, and then face the challenge of actually fixing what was found, often without the specialized expertise, bandwidth, or strategic context needed to prioritize and execute effectively. Findings sit in a backlog. Product roadmaps compress. The same issues reappear in the next assessment.

Sidekick's vSec subscription service breaks this cycle by embedding Sidekick engineers and security leaders directly with the customer's team immediately after an engagement. vSec is not staff augmentation, it's a strategic partnership that brings the same expertise that found the vulnerabilities to the work of closing them.

What vSec Delivers

Hands-On Remediation

Sidekick engineers work side-by-side with your team to implement fixes; from JWT architecture changes and guardrail API deployment to SIEM detection rule tuning and custom agent governance frameworks. We don't just advise; we execute.

Security Program Development

Our leadership team partners with your CISO and product security leadership to build and mature your AI security program; developing policies, governance frameworks, LLM risk management processes, and board-level reporting that demonstrates measurable improvement.

Continuous Validation

As remediations are implemented, Sidekick validates their effectiveness through targeted retesting, confirming that fixes actually work under adversarial conditions rather than just checking a compliance box. AI architectures evolve rapidly; so do the attacks against them.

Knowledge Transfer

Every vSec engagement includes structured knowledge transfer to ensure your team develops the internal capability to maintain and extend security improvements after the engagement concludes. We build capacity, not dependency.

vSec Example Engagement Model

PHASE 1	PHASE 2	PHASE 3	PHASE 4
0–30 Days	**31–90 Days**	**91–180 Days**	**Ongoing**
Critical remediation triage. Deploy prompt injection detection. Begin behavioral logging baselines. Establish session monitoring alerting.	Guardrail API deployment. JWT downscoping architecture. DLP inspection on agent responses. Cross-service SIEM correlation.	JWT revocation capability. Custom agent governance framework. Automated SOAR response playbooks. Risk-based session expiration.	Continuous partnership. Emerging AI threat advisories. Retesting and validation. Regulatory examination preparation.

THE VSEC DIFFERENCE

Traditional consulting engagements end with a report. Sidekick's vSec model starts where the report ends. By embedding the same experts who conducted the assessment directly into remediation, we eliminate the knowledge loss, context switching, and re-explanation that plague traditional consulting handoffs. The engineers who identified the JWT revocation gap are the same ones who architect and implement the fix.

BUSINESS IMPACT & RESULTS

The combined findings from full-spectrum assessment demonstrate the transformative value of evaluating security holistically across every domain. Organizations that act on these results move from a posture of assumed security to one of validated, measurable resilience with clear paths to continuous improvement through the vSec partnership model.

BEFORE SIDEKICK

- AI agent architecture not assessed for post-compromise blast radius
- Prompt injection attack paths unknown and unmonitored
- JWT non-revocability gap unidentified: 60-minute uncontrollable exposure window
- Full-scope JWT credentials issued for every agent tool call
- Session persistence risks not understood: indefinite data retention post-attack
- Custom agent governance gap unaddressed: no validation workflows
- Compliance gaps unknown with no cross-framework mapping
- No behavioral monitoring baselines or AI-specific detection rules

AFTER SIDEKICK

- Discovered attack paths fully documented with blast radius quantification per path
- Prompt injection detection roadmap with phased deployment and content scanning guidance
- JWT revocation and downscoping requirements defined with architectural guidance
- Per-tool permission scoping framework designed and prioritized for roadmap integration
- Session lifecycle risks addressed with expiration policies and sensitivity-based monitoring
- Custom agent governance framework with instruction validation and approval workflows
- Full regulatory compliance mapping across ISO 27001, NIST CSF, and SOC 2
- Deployable detection rules and three-phased detection and response strategy

Client Perspective

> *“We went into this knowing that prompt injection was going to be a problem, we had to assume it was. Really understanding the “what next” issues helped our team thoughtfully think through the controls that made sense for our platform. It was really important to not just receive bland generic recommendations, since user experience is paramount to us.”*
>
> **— Head of Product Security, SaaS Company**

WHY SIDEKICK SECURITY?

Sidekick Security delivers a fundamentally different approach to security assessment, one designed by a former CISO and Red Teamer, built on full-spectrum adversarial methodology, cross-domain intelligence, deployable outcomes, and a continuous partnership model that ensures findings become fixes.

AI-Specific Threat Modeling

We assess AI agent architectures using the same adversarial methodology we apply to traditional infrastructure, mapping trust boundaries, permission inheritance, data flows, and session lifecycles to identify compound attack paths that architecture reviews alone cannot reveal.

Post-Compromise Impact Quantification

For every identified attack path, we quantify the blast radius: data exposed, services accessible, detection likelihood, and containment effectiveness. Organizations get a clear picture of real-world risk, not just a list of findings with severity ratings.

Deployable Remediations

Every engagement delivers more than a report. SIEM rules, WAF configurations, IR playbooks, compliance mappings, and hardening guides, ready to deploy immediately. We close the gap between “findings” and “fixed.”

vSec Partnership Model

Our vSec subscription service embeds Sidekick engineers directly with your team after each engagement. The same experts who conducted the threat model help implement remediations, eliminating knowledge loss and accelerating the path from findings to fixes.

Emerging AI Security Expertise

The AI security threat landscape is advancing rapidly. Sidekick maintains active expertise in prompt injection, indirect injection and agentic attack techniques, ensuring assessments reflect the current state of adversarial capabilities, not last year's research.

Cross-Engagement Intelligence

By assessing AI agent architectures across multiple organizations and platforms, Sidekick brings cross-industry trend insights that help clients benchmark their AI security posture and prioritize investments based on patterns observed across the broader enterprise technology landscape.

READY TO SEE YOUR FULL ATTACK SURFACE?

Contact Sidekick Security to discuss how our full-spectrum assessment methodology and vSec partnership model can strengthen your organization’s security posture, close gaps faster, and build lasting program maturity.

hello@sidekicksecurity.io | sidekicksecurity.io