



AI CYBER ALLIANCE

Welcome to Harness wars.

Follow a tour of real-world exploits across software, containers, marketplaces, and now the hardware itself, revealing why decoupling the AI harness and scaffolding from the model has become a critical security strategy.

<https://ai-cyber-alliance.org/>

Alissa Knight

CEO
Assail



John Van Lowe

Implementation Engineer
DeepTempo





Vigil SOC: Enhancing the SOC

**Agent WorkFlows, Air Gapped Deployments,
MCP RC Updates & Blue Agents Going Purple**

John Van Lowe - AI Cyber Alliance Boston - July 2026

Presenter – whoami – John Van Lowe (JVL)

Professional Background

Aerospace (NASA & US Air Force)

Semiconductor Manufacturing (Micron)

Hosting & Infrastructure (Hostgator, Codero, Patmos)

CyberSecurity & Machine Speed Intelligence (DeepTempo.ai)

Personal Background

From Austin, Texas. Husband and father of twins.

Community founder for the Festival Beach Food Forest,
backyard chicken farmer with 42 chickens, food runner for
Keep Austin Fed.

Group Associations

Austin OWASP, i2Coalition & Secure Hosting Alliance, Austin
ISC2, AI MUG, ATX Javascript, ATX GoLang, AI Camp, VetSec



Intent of the AI Cyber Alliance

Built by Practitioners for Practitioners

The AI Cyber Alliance exists to establish and share real technical depth, open standards, and practical defenses in the generative AI era.

I'm grateful to the local builders, security architects, and organizers who help make these events happen. Interested in getting involved? Let's chat afterword.

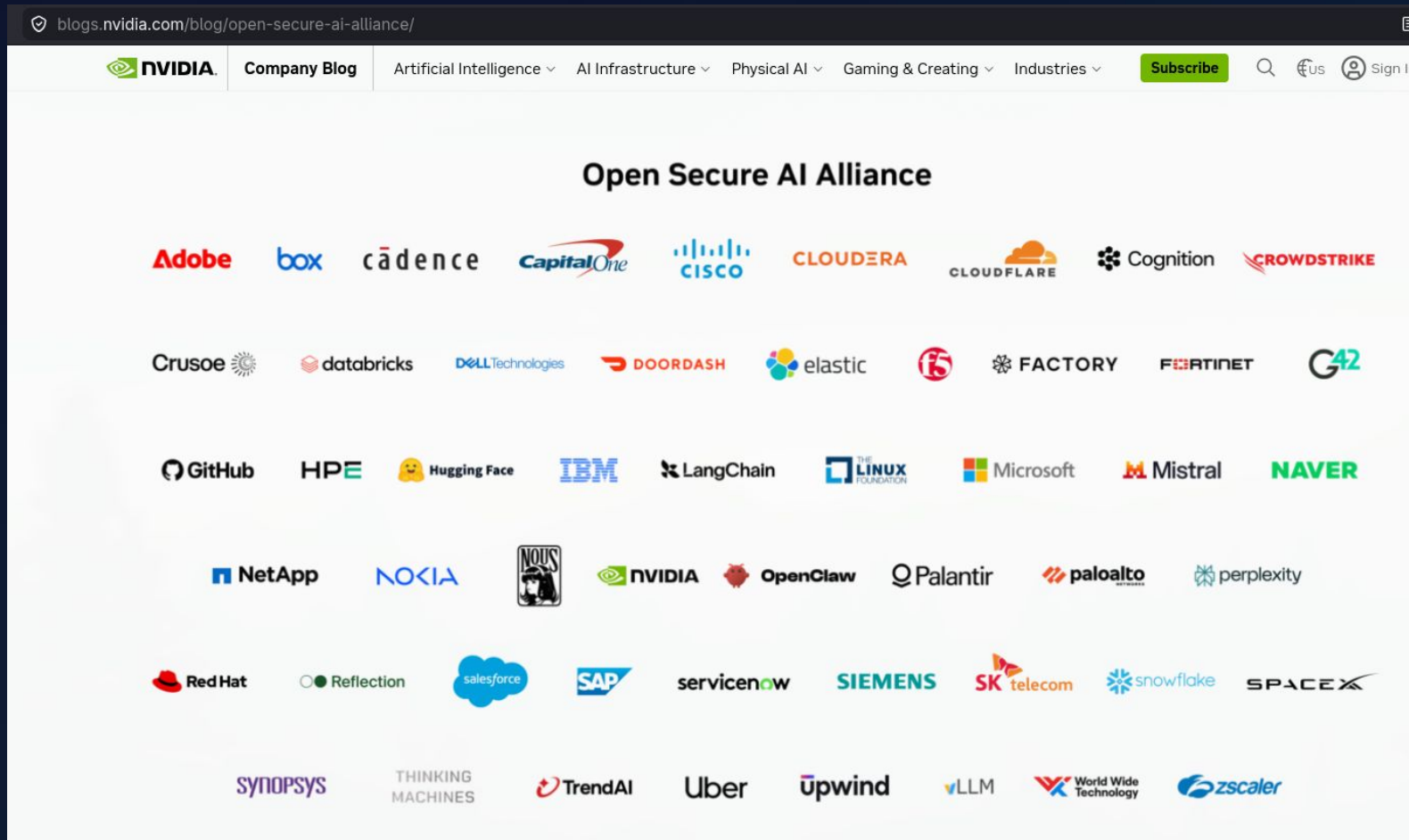
Special thanks to Venture Guides,
DeepTempo, & Assail for direct support!
Thank you to JPM for sponsoring food/drinks.

**Boston, Austin,
SF/Bay, +**



AI CYBER ALLIANCE

Recent News: Open Secure AI Alliance



Legislation or Operator Driven?

What is scope?

Where is there overlap?

OpenAI, Google, and Anthropic not currently in the Open Secure AI Alliance.

Driven to stop an AI Kill Switch?

What are the incentives to increase tech cybersecurity skills in workforce?

Introducing Vigil SOC.org

Skills Driven

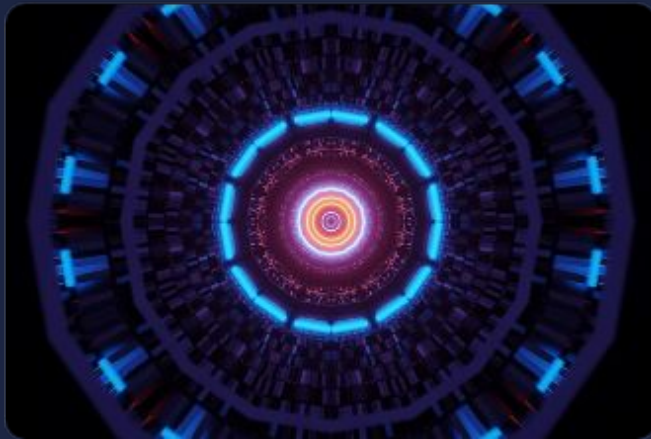
Granular, executable SecOps modules allowing autonomous agents to triage alerts, query SIEMs, and execute precise response playbooks.

MCP Native

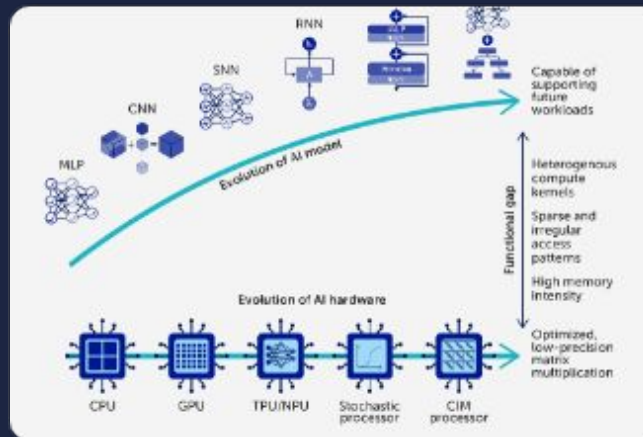
Built around the Model Context Protocol (MCP) to standardize how agents interact with external tools and security datasets.

Apache 2.0 Licensed

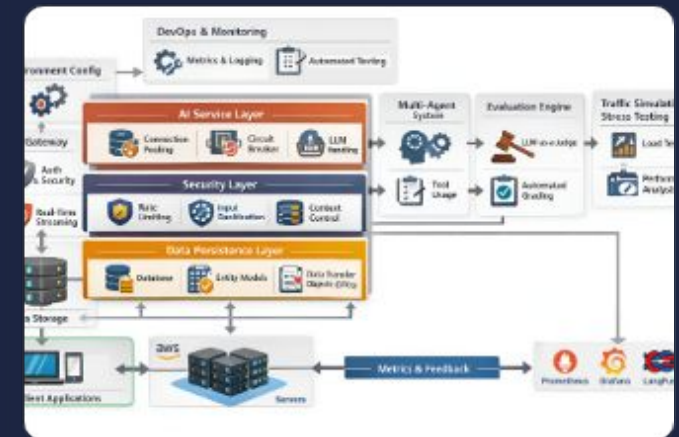
100% open-source foundation ensuring total transparency, vendor independence, and unrestricted enterprise customization.



Bifrost (MCP Gateway)



Mempalace (Local Memory)



Core Harness (Claude SDK)

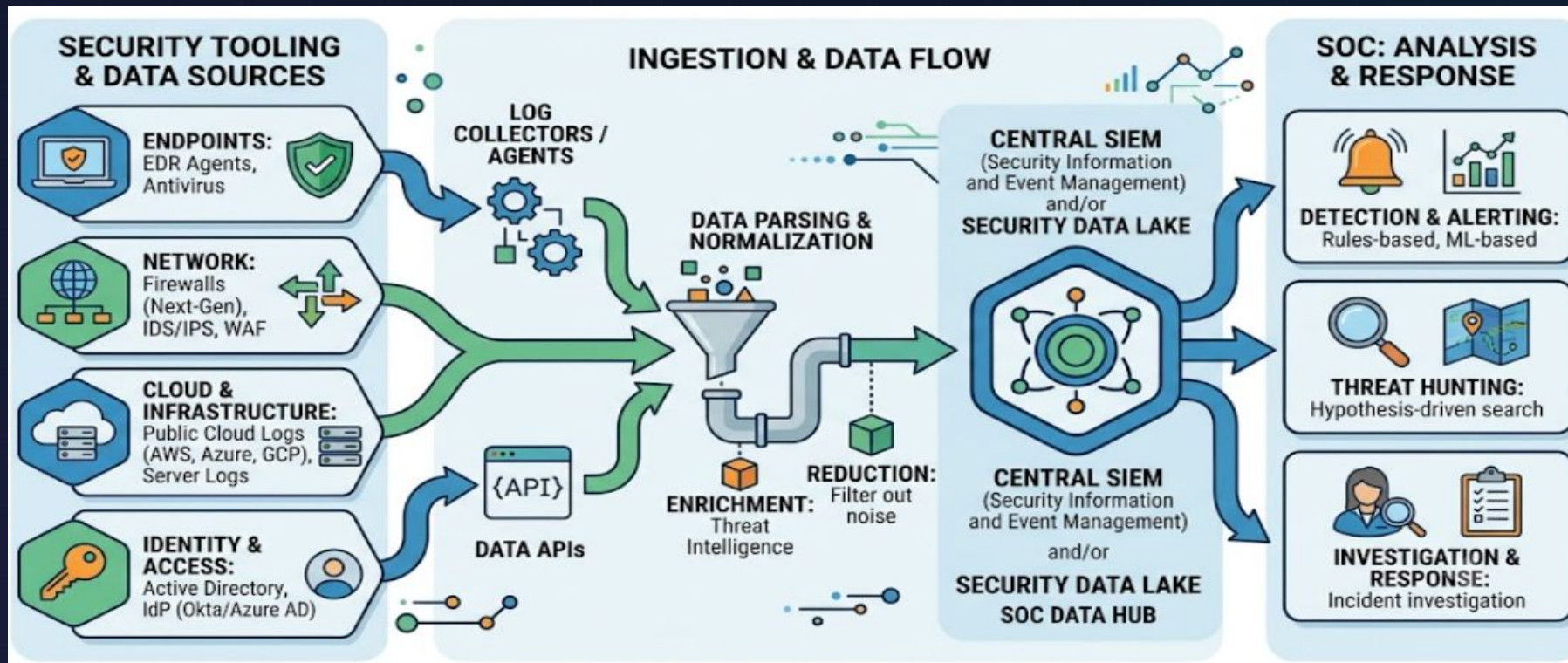
Addressing Modern SOC Challenges

Alert Fatigue & Volume

SOC analysts process thousands of daily telemetry alerts across SIEM, EDR, and Cloud platforms. Up to 70% are false positives, leading to critical burnout and missed threats.

Context Switching Friction

Investigating a single alert requires toggling between 8+ disparate tools for IP lookups, log correlation, and threat intelligence. MTTD and MTTR spike as context is fragmented.

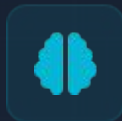


Core Pillars of AI SOC Workflows



1. Ingestion & Triage

Automated parsing and prioritization of high volume SIEM alerts using lightweight microservices and natural language classification.



2. Agent Enrichment

Autonomous lookup of VirusTotal, WHOIS, and endpoint state to produce clear, structured incident summaries for Tier 1 analysts.



3. Governed Action

Enforce human-IN/ON-the-loop approvals for sensitive remediation steps like host isolation or user account suspension.

85%

MTTR REDUCTION

Goal!? Accelerating SOC Outcomes

Automating Tier 1 triage and context gathering slashes Mean Time to Respond from hours to minutes. Analysts are liberated to focus on proactive threat hunting and continuous defense engineering.

Example Vigil SOC Workflows



Cloud Incident



Investigate and respond to cloud security incidents across AWS, Azure, and GCP. Covers identity blast-radius, IAM/role analysis, control-plane vs data-plane attacks, cross-account/cross-tenant pivots, and provider-aware containment.

AGENT SEQUENCE

Investigator

 >

Correlator

 >

MITRE Analyst

 >

Responder

 >

Reporter

EXAMPLE COMMANDS

Run cloud incident response on finding f-20260215...

Investigate this suspicious IAM activity in AWS

Cloud incident: unauthorized S3 access from exter...

Respond to Azure AD credential compromise alert

Run cloud-incident workflow for this GCP SCC find...

 History

▶ Run workflow



Forensic Analysis



Post-incident digital forensics with evidence preservation, malware deep-dive, network forensics, and audit-ready documentation suitable for legal proceedings.

AGENT SEQUENCE

Forensics

 >

Malware Analyst

 >

Network Analyst

 >

Reporter

EXAMPLE COMMANDS

Conduct forensic analysis on the compromised host...

Post-incident forensics for case CASE-20260215-br...

Forensic examination of these artifacts for legal...

Run forensics workflow on this data breach incide...

 History

▶ Run workflow



Full Investigation



Comprehensive multi-agent investigation with MITRE ATT&CK mapping, cross-signal correlation, response planning, and detailed reporting.

AGENT SEQUENCE

Investigator

 >

MITRE Analyst

 >

Correlator

 >

Responder

 >

Reporter

EXAMPLE COMMANDS

Fully investigate this finding and all related ac...

Do a complete investigation of case CASE-20260215...

Deep dive into these suspicious lateral movement ...

Run full investigation workflow on this cluster o...

 History

▶ Run workflow

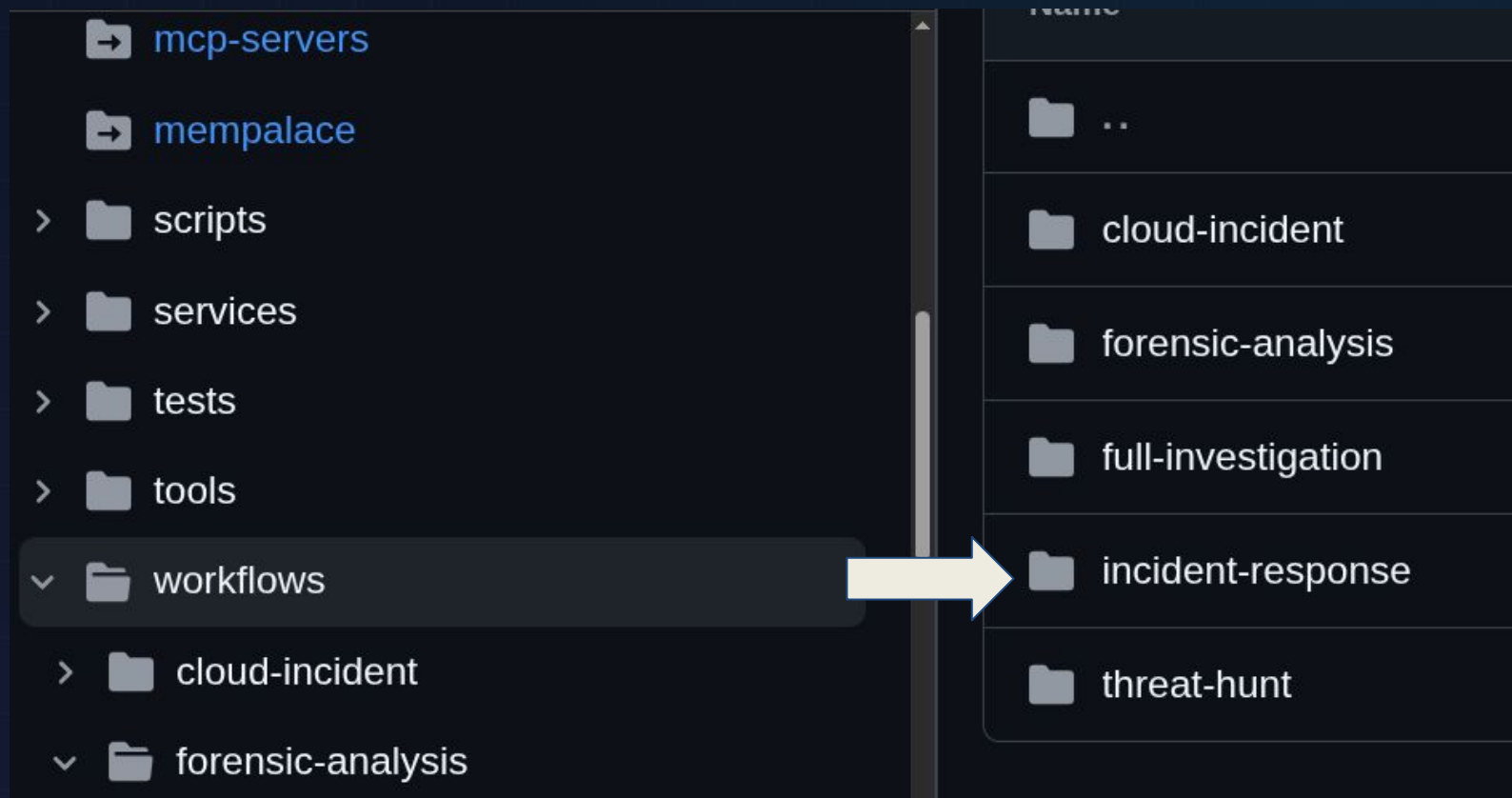
Workflows: Incident Response Overview

Beyond Static Playbooks

Traditional playbooks rely on hardcoded conditional rules that fail when attack patterns evolve.

AI workflows combine deterministic code execution with dynamic LLM agent reasoning. It allows us to evaluate incoming telemetry in context before taking action.

Distributed in Markdown, revision controlled within Git.



<https://github.com/Vigil-SOC/vigil/blob/main/workflows/incident-response/WORKFLOW.md>

Incident Response Workflow In Depth

Incident Response Workflow

Multi-agent incident response workflow following the NIST Incident Response framework. Sequences four specialized agents to rapidly triage, deeply investigate, contain threats, and produce audit-ready documentation.

USE CASES:

- Security alert has fired requiring immediate review
- Finding is flagged as critical
- An active threat (malware, C2, data exfiltration) is occurring
- You need end to end triaging for a report

name	incident-response					
description	Respond to active security incidents with rapid triage, deep investigation, containment, and documentation. Follows NIST					
agents	triage	investigator	responder	reporter		
tools-used	get_finding	list_findings	nearest_neighbors	search_detections	create_approval_action	update_case
use-case	Active incident response -- an alert fires and the SOC needs to triage, investigate, contain, and document.					
trigger-examples	Run incident response on finding f-20260215-abc123			We have an active incident -- ransomware detected on HOST		

Incident Response Agent Phases

Phase 1: Triage & Classify (Triage Agent)

Purpose: Rapid assessment of the alert -- severity scoring, false-positive check, categorization.

Tools: `get_finding`, `list_findings`

Output: Severity score, alert category, priority level, escalation decision

Short-circuit: If classified as False Positive, skip to Phase 4 for a brief dismissal report.

Phase 2: Deep Investigation (Investigator Agent)

Purpose: Root cause analysis, evidence collection, timeline reconstruction, cross-source correlation.

Tools: `get_finding`, `list_findings`, `nearest_neighbors`, `search_detections`

Output: Root cause, attack vector, affected entities (IPs/hosts/users), evidence chain, related findings, timeline

Phase 3: Contain & Respond (Responder Agent)

Purpose: NIST IR containment -- isolate hosts, block IPs, revoke credentials, plan remediation.

Tools: `create_approval_action`, `get_finding`, `update_case`

Output: Containment actions with confidence scores, approval requests, remediation checklist, blast radius assessment

Phase 4: Document & Report (Reporter Agent)

Purpose: Generate audience-tailored incident report with executive summary, technical details, and lessons learned.

Tools: `get_case`, `list_findings`, `create_attack_layer`

Output: Complete incident report, MITRE ATT&CK layer, event timeline, recommendations

① Vigil Assistant ⌚ 💡 📄 ⚙️ 🗑️ ✕

③ **Responder** — Containment **recommended, not auto-executed** (exact ARNs/principals unverifiable without evidence; nothing submitted to approval queue): revert trust policy, rotate/disable service-account keys, re-enroll admin MFA, run IAM Access Analyzer for blast radius.

④ **Reporter** — Detection-coverage check revealed a **blind spot**: T1484.002 (Trust Modification) has only **1** detection and **no AWS-native trust-policy-tampering coverage** — the exact technique used. (Coverage elsewhere: T1078.004=42, T1098.003=7, T1531=9, T1556=16.)

Required next actions (to unblock)

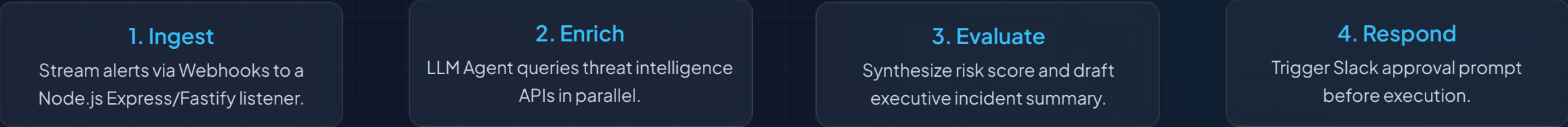
1. Ingest CloudTrail: `UpdateAssumeRolePolicy` + MFA-delete calls with actor identity, sourceIP, userAgent, timestamps.
2. Run IAM Access Analyzer to enumerate the backdoor's blast radius.
3. Pull auth/session logs to determine **how** the service account was compromised.

Memory persisted

- Investigation drawer + KG relationships
(`service_account` → `modified_trust_policy_of` → `Production-Admin` ; → `deleted_mfa_device_of` → `key admin account`) + agent diary entry.

1,245 / 200k tokens ~\$0.0187-\$0.3259 · heuristic

Example Workflow: Vigil SOC Incident Response



Human-in-the-Loop & HOTEL Safety



What is the remediation?
Can it be quickly reverted?
Can a case wait?

What are the self-demotion events?

Deterministic Guardrails & Confidence Scores

AI can make confidence gated high-impact non-destructive actions autonomously. As models and tooling mature it is becoming more common to allow autonomy with potentially destructive actions. In some cases we are using evaluations from subsequent agent using a different model.

$$C_{\text{total}} = \left(\frac{\sum_{i=1}^n w_i \cdot c_i}{\sum_{i=1}^n w_i} \right) \times (1 - P_{\text{fp}})$$

Where:

- $c_i \in [0.0, 1.0]$: Confidence score of evidence factor i (e.g., LLM analysis confidence, threat intel indicator match score, telemetry match score).
- w_i : Assigned weight for evidence component i .
- $P_{\text{fp}} \in [0.0, 1.0]$: Penalty factor based on historical false-positive rates for the rule/detector.

Work ultimately led to open source publishing of another internal tool: SOC Bench.org

Inference Pivot: Air-Gapped & Local AI Trends

Strict Isolation

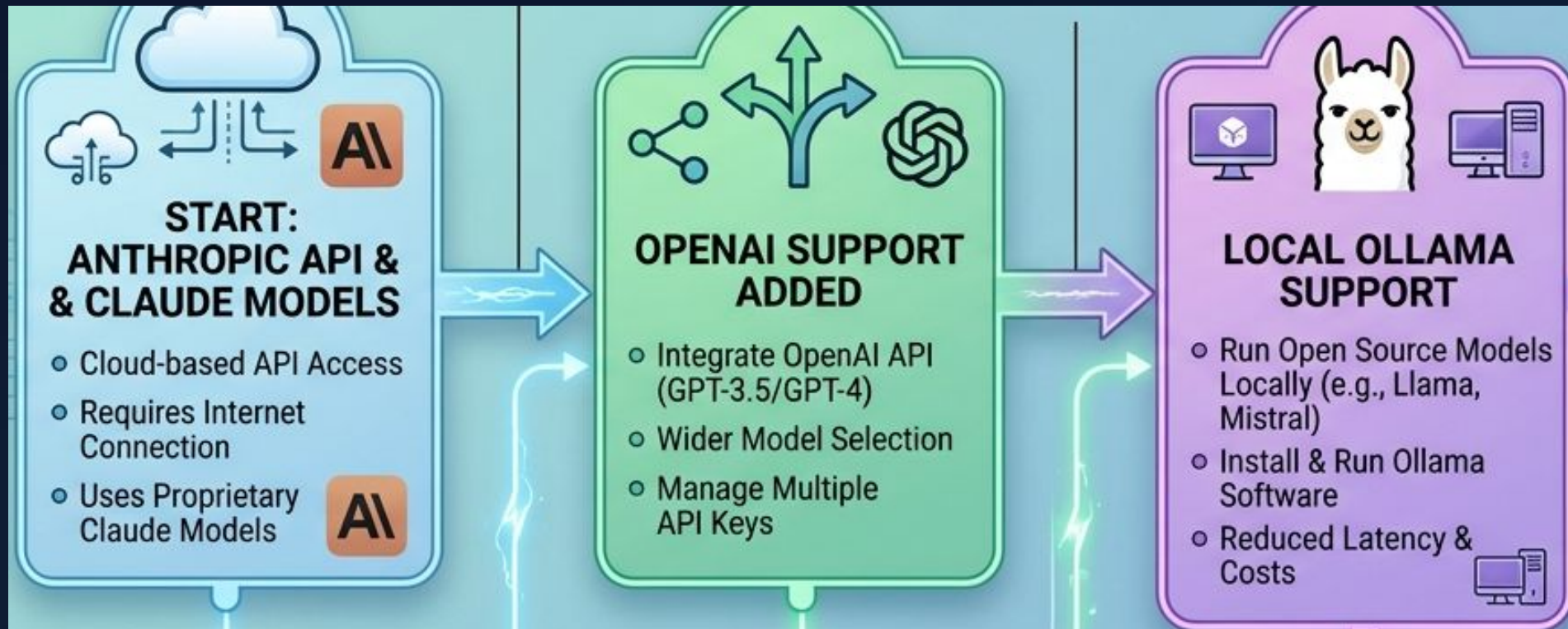
Regulated industries (defense, finance, healthcare) demand strict air-gapped environments where third-party API egress is completely prohibited.

Zero Telemetry Leaks

Local inference guarantees sensitive contextual data and proprietary codebase logic never leak into external model training sets or remote logs.

Offline Resiliency

Autonomous security agent pipelines must remain fully functional during network partitioning or adversarial internet degradation.



Self Hosted Local Inference

Ollama Runtime

Ubiquitous lightweight local runner. Ideal for developer desktop testing, edge nodes, and rapid prototyping of HG F-quantized models.

Provides standardized API endpoints mimicking OpenAI structures for easy local drop-in capability.

Lemonade Server (AMD)

Purpose-built server engine designed for optimized hardware utilization across AMD Ryzen AI, ROCm GPUs, and NPU architectures.

Delivers high-throughput local inference with multi-engine backend flexibility (llama.cpp, FastFlow) for enterprise workstations.

Providers

Model Assignment

Operations

Budgets

Model Assignment

Pick a provider + model for each system component. Unassigned rows fall back to the chat_default assignment. The model list is live-queried from each provider.

Component	Provider	Model
Chat (Default) Fallback for interactive chat and every component below when unset.	anthropic	Claude Opus 4.8
Triage Agent Automated alert triage — cheaper/faster models work well here.	gh200-on-prem	qwen3.5:latest
Investigation Agents Investigator, Threat Hunter, Correlator, etc. — the heavy lifters.	gh200-on-prem	gpt-oss:120b
Orchestrator — Planning Generates the investigation plan from the initial finding.	gh200-phoenix	hermes3:70b
Orchestrator — Review Reviews and approves sub-agent output at the end of an investigation	anthropic	Claude Opus 5
Context Summarization Compresses long conversations — a cheap model is usually fine.	gh200-on-prem	gpt-oss:120b
Report Generation Reporter agent output — clarity and structure matter more than depth	gh200-on-prem	llama3.3:latest

Semantic Inference Routing Protocol

SIRP

~~Security Incident Response Plan~~
Semantic Inference Routing Protocol

Why SIRP Matters for Sensitive Workloads

SIRP dynamically evaluates prompt intent, sensitivity classification, and payload policy prior to model execution.

SIRP factors in data attributes of category and sensitivity and process attributes for task complexity.

It intelligently routes based on Cost Optimization, Urgency Prioritization, Privacy-Aware Handling.

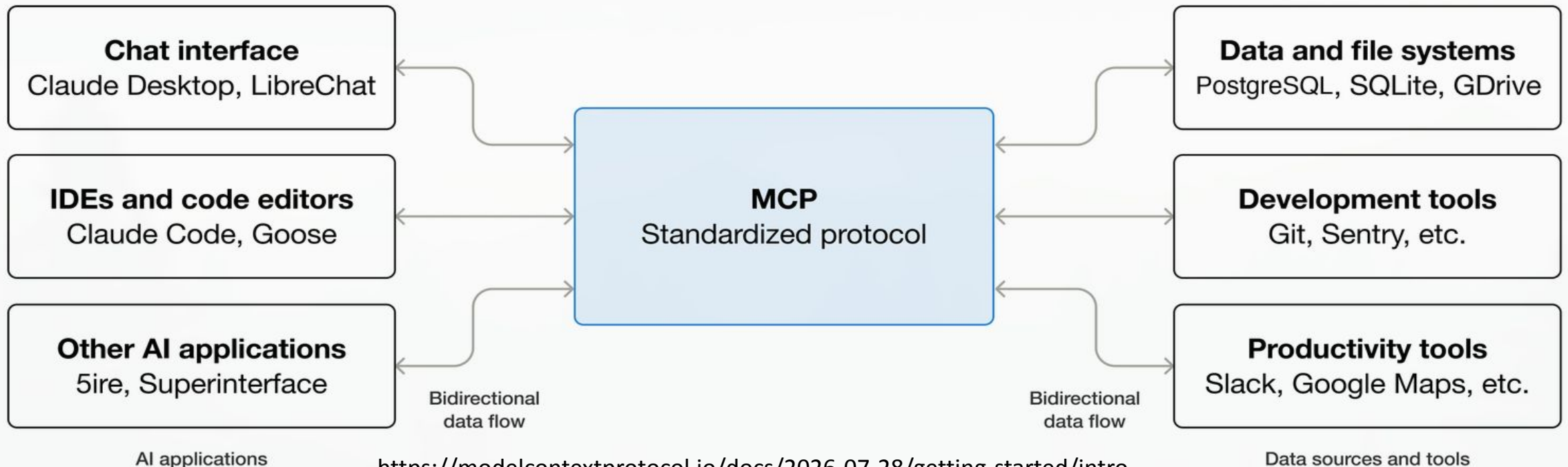


Model Context Protocol Updates

MCP RC

Protocol Release History

Released by Anthropic in November 2024, four dated releases prior to current RC 2026-07-28 (sometimes referred to as MCP 2.0).



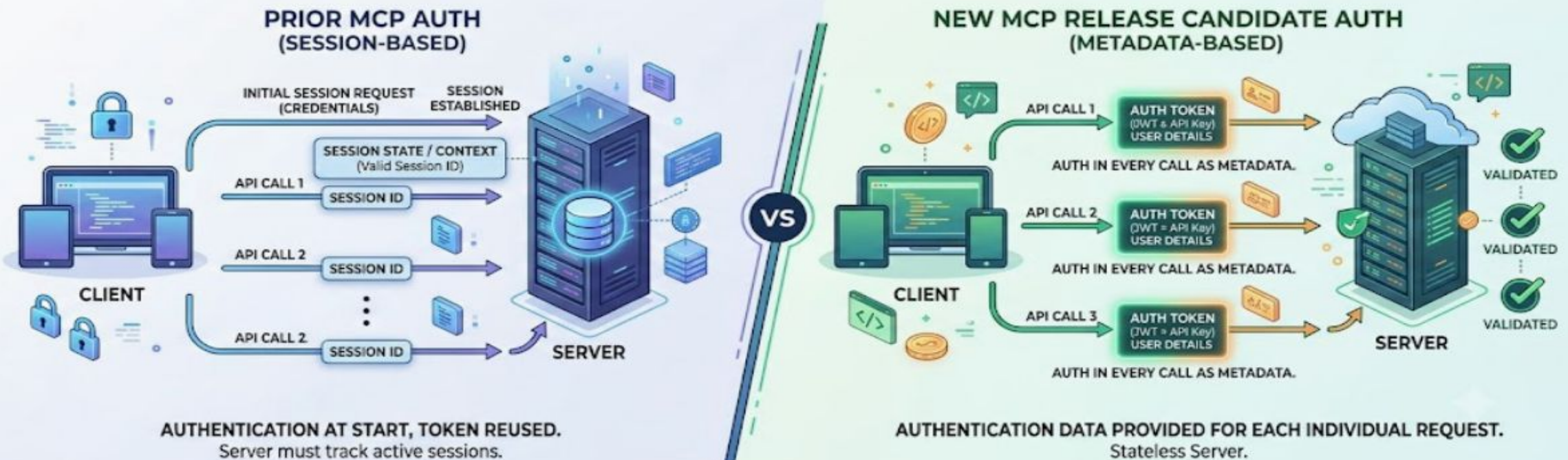
Model Context Protocol Authentication

MCP Auth

Authorization Updates

Prior every session started with auth. Now every request has authentication metadata.

MCP AUTHENTICATION COMPARISON: SESSION VS. PER-CALL

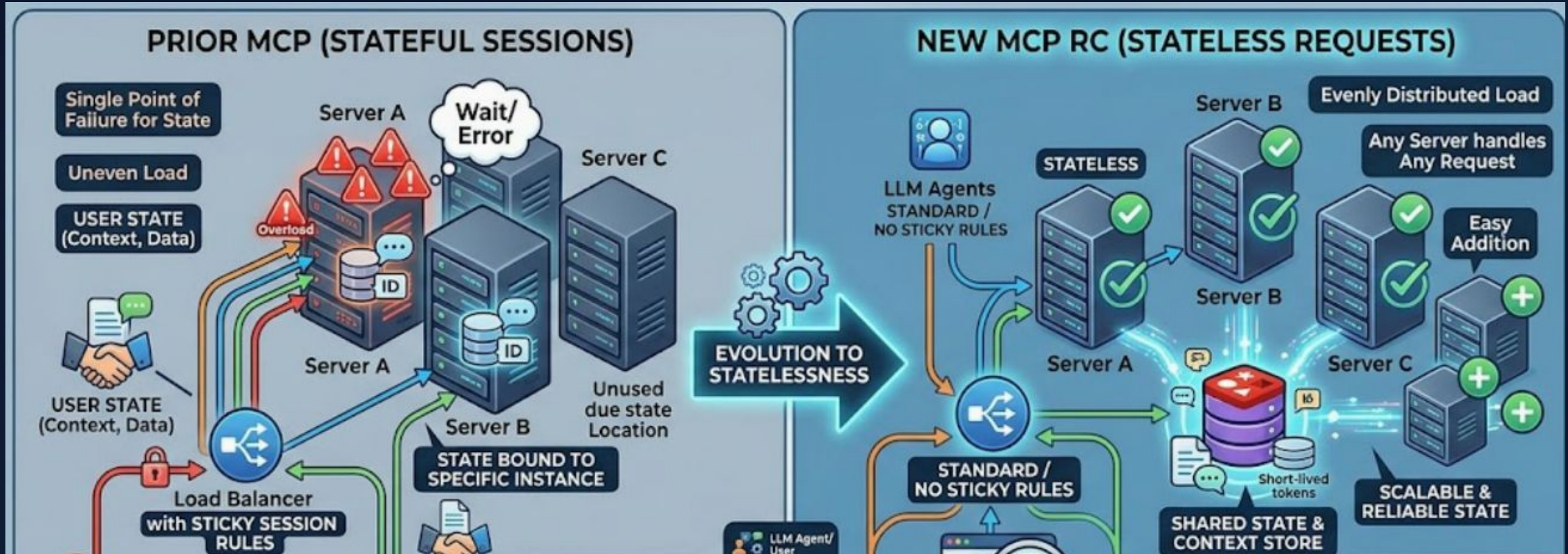


Model Context Protocol Scaling

MCP Scale

MCP Scaling

The protocol is stateless and allows deployment without sticky sessions (for prior Auth). No more session persistence via load balancers and native support for K8S ingress without customization.



Model Context Protocol Apps & Tasks

MCP Misc

Apps, Tasks, and Governance

MCP Apps allow the models to return more than text, MCP Tasks permit long running tasks that require more than a single tool call or inference run, and 12 month deprecation policy gives assurances.

NEW MCP RELEASE CANDIDATE FEATURES

MCP APPS

APPLICATION LIFECYCLE MANAGEMENT

- 1 Streamlined App Onboarding**
Simplified wizards for app creation and deployment
- 2 Cross-Cluster Deployment**
Deploy and manage applications across multiple clusters
- 3 Enhanced Observability**
Integrated application health dashboards and performance metrics
- 4 Centralized Logging**
Aggregated logs and monitoring for all apps

MCP TASKS

WORKFLOW AUTOMATION & SCHEDULING

- 1 Workflow Orchestration (DAGs)**
Support for complex multi-step task dependencies
- 2 Event-Driven Scheduling**
Automate tasks based on triggers, events, or time
- 3 Granular Resource Control**
Optimize resource usage for job execution
- 4 Task Execution History**
Detailed logs, status tracking, and retry logic

GOVERNANCE

POLICIES & COMPLIANCE

- 1 Policy Guardrails**
Automated enforcement of constraints and security rules
- 2 Audit Trails & Compliance**
Comprehensive logging of actions and changes
- 3 Granular Access Control (RBAC)**
Fine-grained role-based permissions
- 4 12-Month Deprecation Policy**
Clear timelines, Cios and guidelines for retiring older versions and features, promoting lifecycle planning

Vigil SOC Blue to Purple?

Community Contributions Reshaping Intent

Vigil SOC was originally positioned as a defensive “blue team” tool. In recent conversations we’ve seen deployments where Vigil SOC is also offensive. There are use cases where MCP tooling for NMap, shodan, and existing third party providers (some black box) are in the workflows.

Red Team: Emulates adversaries to find exploitable weaknesses.

Blue Team: Operates defenses to detect and respond to attacks.

Purple Team: Coordinates collaboration to measurably improve both.

White Team: Deconfliction? The actual human? Another future talk about a new agent?



Vigil SOC Path to Purple?



// red ops · recon → exploit

blue ops · triage → report //

01

recon

subdomain & port
discovery

02

exploit

CVE chaining,
privesc, lateral

03

triage

scoring, FP filtering,
escalation

04

investigate

timelines, IOC,
ATT&CK

05

respond

isolate, disable,
blocklist (gated)

06

report

audit writeups,
detections

● red: autonomy is the feature

● blue: humans stay in the loop, on purpose

Interested in contributing? Accepting PR's for both feature/fix and blog content related to Vigil SOC deployments and best practices.

Questions?

Thank you for listening! Connect and find more:



AI CYBER ALLIANCE

• COMMUNITY-LED • OPEN SOURCE • VENDOR-NEUTRAL

Making cyber faster and more intelligent — together, in the open.

Sharing hands-on experiences using AI, software, and dedicated humans. Two technical talks every two weeks, across the US. Built by the community – no product pitches.

aicyberalliance.org

retroencabulation.com

[Vigil SOC.org](https://VigilSOC.org)

socbench.org

linkedin.com/in/johnvl