MCP-Enabled Intelligent Cyber Agent for Red and Blue Team Automation

Srujan Dasari, M.S in Cybersecurity

FACULTY MENTOR: Daniel Galeon, MBA

Katz School of Science and Health

Introduction

The problem:
- Pentesters lose hours on manual recon and enumeration before real analysis begins.
- SOC analysts face alert fatigue and miss subtle IOCs buried in log noise.
- LLMs like ChatGPT can explain but cannot act on live systems.

Our approach:
Give the LLM "hands" via the **Model Context Protocol (MCP)** turning a passive advisor into a controlled, auditable operator across red and blue team workflows.

Research Question:
Can a single AI agent bridging offensive and defensive security operations through the Model Context Protocol safely automate the "Tier 1" labor of modern SOC environments?

Approach

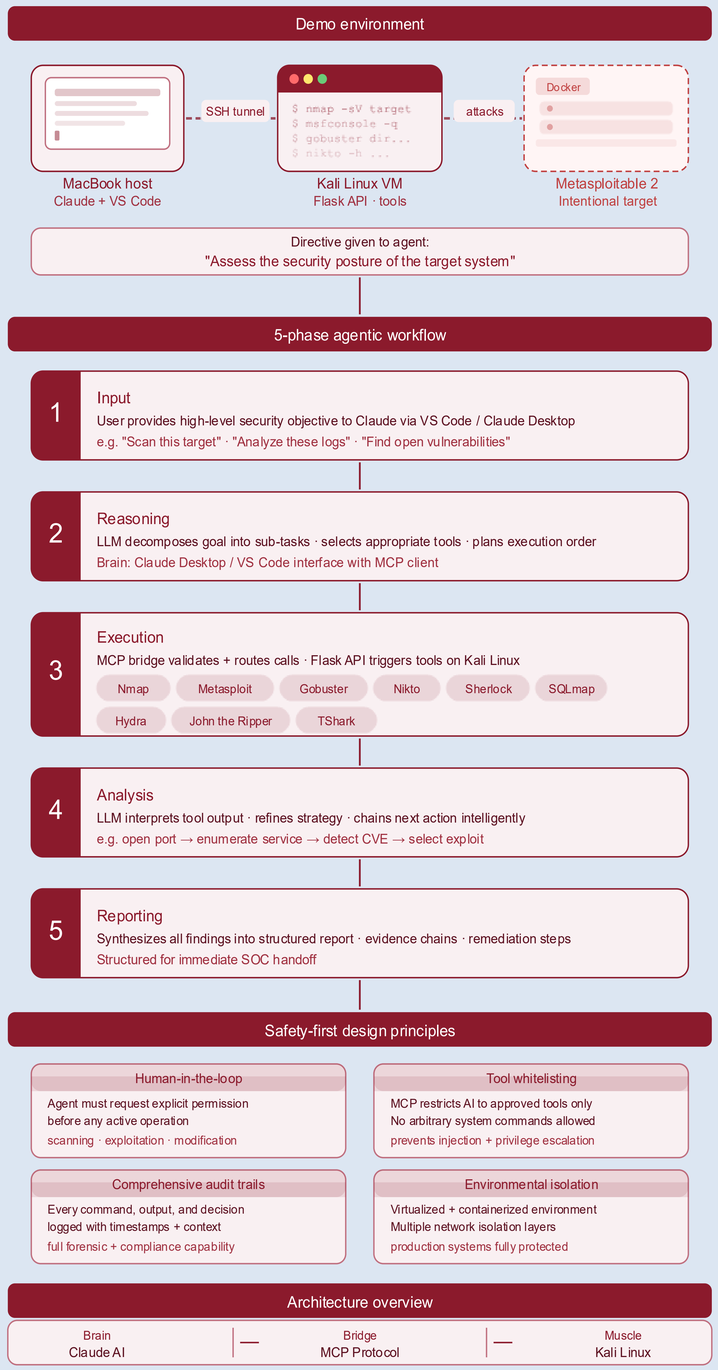
To answer the research question, we built a **five-phase agentic workflow** that mirrors how a Tier 1 SOC analyst operates:

- **Input —** high-level goal (e.g., "assess this target").
- **Reasoning —** LLM decomposes goal into sub-tasks.
- **Execution —** runs Kali tools via MCP-validated calls.
- **Analysis —** interprets output, picks next step (port → service → exploit).
- **Reporting —** structured findings, ready for SOC handoff.

Architecture: Brain → Bridge → Muscle

- **Brain:** Claude / VS Code (Copilot) interprets intent.
- **Bridge:** MCP enforces safety whitelisting and audit logs answer the "safely" in our research question.
- **Muscle:** Kali Linux Nmap, Metasploit, Nikto (red); TShark, log parsers (blue).

Workflow



Results

Case Study: Metasploitable 2 (Red Team)
Directive: "Assess the security posture of the target."
- Agent autonomously chained **Target IP** → **Nmap** → **Metasploit** with no manual intervention.
- Identified **17 open ports** and flagged vsftpd 2.3.4 backdoor (CVE-2011-2523) for human approval.
- End-to-end recon to report time: **~4 min** vs. ~45 min manual baseline (~91% reduction).
- Every action logged with timestamp; permission gate halted exploit until approved.

Defensive Demo (Blue Team)
Directive: "Analyze these auth logs for compromise."
- Parsed **10K+ log lines**; surfaced an SSH brute-force spike (200+ failedlogins from one IP).
- TShark caught port-scan and DNS-tunneling signatures missed bybaseline rules.
- Generated structured incident report with evidence chain &remediation.
- Triage time vs. manual review: **~3 min vs. ~30 min**.
- **Safety check:** 0 unauthorized commands across all test runs whitelistheld.

Conclusions

Answer to the research question: Yes a single MCP-enabled agent can **safely automate Tier 1 SOC labor** across both red and blue team workflows.
- **Single agent, both sides:** one workflow handled offensive recon and defensive log triage breaking the red/blue silo.
- **Safely:** localhost binding, SSH tunneling, tool whitelisting, and audit logs 0 unauthorized commands in testing.
- **Tier 1 automated:** ~90% time reduction on recon and log-triage tasks vs. manual baseline.

Limitations:
- Lab-only testing; enterprise networks will vary.
- Quality bound by LLM reasoning on ambiguous goals.
- Single-user only; no concurrent access yet.

Future Work:
- Live SIEM integration (Splunk, Elastic).
- Auto-remediation: patches, firewall rules, hardening.
- Multi-stage chains: recon → exploit → privesc → persistence.

References

- Anthropic. (2024). *Model Context Protocol specification*. https://modelcontextprotocol.io
- MITRE. (2024). *ATT&CK Framework for Enterprise*. https://attack.mitre.org
- NIST. (2018). *Cybersecurity Framework v1.1*. National Institute of Standards and Technology.
- Shostack, A. (2014). *Threat Modeling: Designing for Security*. Wiley.
- Lyon, G. (2009). *Nmap Network Scanning*. Nmap Project.