BAGLEY: Orquestração Automatizada Baseada em LLM de Ciclos de Purple Teaming e MITRE CALDERA

Junior Petry Machado[1], **Alexandre Alyson Sousa Pontes**[1], **Ramon dos Reis Fontes**[1]

[1] Instituto Metrópole Digital (IMD) – Universidade Federal do Rio Grande do Norte (UFRN)
Natal – RN – Brasil

{junior.petry.115, alexandre.pontes.101, ramon.fontes}@ufrn.edu.br

Abstract. *The rapid evolution of cyber threats necessitates dynamic defense strategies that move beyond rigid models and infrequent security testing. This paper presents BAGLEY, an integrated orchestration platform that automates the Purple Teaming cycle by linking MITRE CALDERA's offensive emulation with Wazuh SIEM's defensive analysis. Orchestrated by a Large Language Model (LLM), the system abstracts technical complexity by translating high-level tactical intentions into executable multi-platform commands. BAGLEY features a unique Cognitive Loop that performs automated telemetry matching, identifies detection gaps, and corrects execution failures in real-time through an autonomous resilience mechanism. Experimental results across multiplatform environments demonstrate the system's ability to execute stealthy Living off the Land (LOTL) attacks; filter benign system noise to isolate high-relevance indicators; and formulate actionable forensic reports with ready-to-deploy mitigation recommendations.*

Resumo. *Para mitigar os riscos da rápida evolução das ameaças cibernéticas, são necessárias estratégias de defesa dinâmicas que superem a eficácia de modelos rígidos e testes esporádicos. Este artigo apresenta o BAGLEY, uma plataforma de orquestração integrada que automatiza o ciclo de Purple Teaming ao vincular a emulação ofensiva do MITRE CALDERA com a análise defensiva do Wazuh SIEM. Orquestrado por um Large Language Model (LLM), o sistema abstrai a complexidade técnica traduzindo intenções táticas de alto nível em comandos multiplataforma executáveis. O BAGLEY apresenta um Loop Cognitivo exclusivo que realiza a correspondência automatizada de telemetria, identifica lacunas de detecção e corrige falhas de execução em tempo real por meio de um mecanismo de resiliência autônoma. Resultados experimentais em ambientes multiplataforma demonstram a capacidade do sistema de executar ataques furtivos do tipo Living off the Land (LOTL); filtrar o ruído do sistema para isolar indicadores de alta relevância; e formular relatórios forenses acionáveis com recomendações de mitigação prontas para implantação.*

1. Introdução

O cenário contemporâneo de cibersegurança é caracterizado pela rápida evolução das ameaças e pelo aumento da complexidade dos ambientes computacionais, exigindo estratégias mais flexíveis para prevenção, detecção e resposta a incidentes

[Tellache et al., 2025]. Nesse contexto, as práticas de *Red Team* e *Blue Team* tornaram-se componentes fundamentais para o fortalecimento da resiliência organizacional.

Enquanto o *Red Team* busca emular adversários reais, explorando vulnerabilidades e avaliando a resistência de uma organização a ataques plausíveis [Hardikar et al., 2025], o *Blue Team* monitora o ambiente, detecta anomalias e responde a incidentes. O *Purple Teaming* integra essas duas perspectivas por meio de uma colaboração contínua entre equipes ofensivas e defensivas, promovendo um ciclo iterativo de execução, análise e aprimoramento das capacidades de detecção e resposta, com o objetivo de reduzir lacunas de visibilidade e fortalecer continuamente a postura de segurança da organização.

No entanto, a crescente complexidade dos ataques cibernéticos e a velocidade com que novas técnicas emergem tornaram os modelos tradicionais de testes de segurança e resposta a incidentes insuficientes [Nguyen et al., 2021]. Consequentemente, as organizações enfrentam um cenário em que falhas de detecção podem permanecer invisíveis por longos períodos, enquanto os atacantes exploram automação e inteligência artificial para ampliar o alcance e a sofisticação de suas campanhas.

Nesse contexto, *frameworks* como o MITRE ATT&CK e plataformas de emulação de adversários, como o CALDERA, representam o estado da arte na simulação de ataques [MITRE, 2025]. Entretanto, essas ferramentas ainda dependem significativamente da atuação de especialistas para definir objetivos, selecionar Táticas, Técnicas e Procedimentos (TTPs) e interpretar os resultados obtidos. A incorporação de *Large Language Models* (LLMs) amplia esse potencial ao permitir a interpretação de cenários complexos, a geração de sequências de ataque plausíveis e a análise contextualizada de grandes volumes de eventos de segurança. Avanços recentes reforçam essa tendência: sistemas baseados em LLM, como o Mythos da Anthropic [Carlini et al., 2026], já demonstram capacidade de identificar e explorar vulnerabilidades *zero-day* de forma autônoma, evidenciando o potencial da inteligência artificial para apoiar operações de *Red Team*.

Esta pesquisa parte da constatação de que muitas organizações ainda realizam avaliações ofensivas apenas por meio de testes de intrusão periódicos, produzindo longos intervalos sem um diagnóstico atualizado da exposição a ameaças [Singer et al., 2025].

Diante desse cenário, este trabalho tem como objetivo propor e validar uma arquitetura capaz de automatizar ciclos de *Purple Teaming* por meio da integração entre modelos de linguagem, a plataforma MITRE CALDERA e um SIEM, permitindo correlacionar continuamente ataques simulados com a telemetria operacional e apoiar o aperfeiçoamento dos mecanismos de detecção e resposta.

Diferentemente de abordagens que apenas automatizam a execução de ataques ou a geração de regras de detecção, a principal contribuição deste trabalho consiste na proposição de um Loop Cognitivo para *Purple Teaming*, capaz de integrar continuamente a emulação de adversários, a análise da telemetria operacional e a geração automatizada de mecanismos corretivos. Essa abordagem estabelece um ciclo de retroalimentação entre capacidades ofensivas e defensivas, permitindo identificar lacunas de detecção, correlacionar evidências observadas durante os ataques e propor melhorias de forma sistemática. Como resultado, se reduz a dependência de intervenção manual durante campanhas de *Purple Teaming* e se estabelece um mecanismo contínuo de evolução da postura defensiva.

2. Fundamentação Teórica

Esta seção fundamenta o BAGLEY em três pilares essenciais: a aplicação de LLMs na cibersegurança, focada em conciliar o raciocínio estratégico com a precisão operacional; os requisitos para a emulação dinâmica de adversários, que permite a transição de ataques estáticos para comportamentos adaptativos; e o *framework* MITRE CALDERA, utilizado como base tecnológica para a orquestração e execução automatizada de táticas ofensivas baseadas no MITRE ATT&CK.

2.1. Large Language Models na Cibersegurança

O rápido avanço dos LLMs mudou seu papel de assistentes de uso geral para agentes especializados capazes de interpretação contextual e análise detalhada em cibersegurança e forense [Yin et al., 2025]. Sua capacidade de organizar informações dispersas, interpretar sinais ambíguos e raciocinar sobre cenários extensos supera as abordagens tradicionais na detecção de ameaças, análise de vulnerabilidades e suporte à resposta a incidentes.

No domínio ofensivo, os LLMs demonstram utilidade prática ao analisar documentação não estruturada, revisar *exploits* de terceiros e interpretar relatórios técnicos para identificar os pré-requisitos e características de várias técnicas de ataque [Wang et al., 2024]. Ao estabelecer ligações lógicas entre estágios que não estão explicitamente descritos nas fontes, esses modelos facilitam a construção de sequências abrangentes de ações de adversários. *Frameworks* como o AURORA [Liu et al., 2018] exploram esse potencial combinando modelos generativos com métodos clássicos de planejamento para produzir fluxos ofensivos estruturados que imitam o comportamento do operador humano.

Apesar de sua proficiência conceitual em estratégias ofensivas, os LLMs enfrentam obstáculos significativos na execução em múltiplos *hosts* devido à sua tendência de gerar comandos inconsistentes ou operacionalmente inválidos. Essa lacuna é impulsionada principalmente por alucinações, onde uma única saída incorreta pode desviar um ataque automatizado ou, mais criticamente, comprometer os esforços de Resposta a Incidentes, levando a decisões falhas que atrasam a contenção e aumentam os danos operacionais.

Para mitigar os riscos, pesquisas recentes têm focado no ajuste fino supervisionado (*fine-tuning*) de modelos menores e na implementação de arquiteturas de Geração Aumentada por Recuperação (RAG), que restringem as saídas a informações previamente validadas. O uso efetivo de LLMs na cibersegurança depende, portanto, de uma integração cuidadosa entre suas capacidades generativas e ferramentas capazes de validar, contextualizar e traduzir com precisão as instruções produzidas em ações técnicas.

2.2. Segurança Ofensiva e Emulação de Adversários

O cenário defensivo moderno é caracterizado pela rápida evolução de ferramentas ofensivas, técnicas e *payloads* customizados [Wang et al., 2024]. Esse fluxo contínuo envolve a descoberta constante de vulnerabilidades e a disseminação de *exploits* funcionais e *kits* de ataque personalizados. Consequentemente, a proteção de infraestruturas complexas exige conjuntos de dados (*datasets*) de alta fidelidade para *benchmarking* e treinamento, fundamentados em estratégias de defesa informadas por ameaças.

Para serem eficazes, os *datasets* ofensivos devem priorizar três requisitos centrais: reprodutibilidade, diversidade e realidade [Wang et al., 2024]. A reprodutibilidade garante que os ataques possam ser fielmente reconstruídos através de documentação detalhada de pré-requisitos, *scripts* e infraestrutura, permitindo que os defensores coletem telemetria em múltiplas camadas, como em chamadas de sistema e tráfego de rede. A diversidade garante uma ampla cobertura em várias superfícies de ataque, enquanto a realidade assegura que os comportamentos emulados estejam alinhados com relatórios documentados de *Cyber Threat Intelligence* (CTI) [Wang et al., 2024].

Na prática, a auditoria de *logs* do sistema é essencial para detectar atividades suspeitas, porém o enorme volume de dados e a falta de rotulagem adequada apresentam desafios analíticos significativos [Huang et al., 2022]. *Frameworks* como o SAGA (S*ynthetic Audit Log Generation for APT Campaigns*) [Huang et al., 2022] abordam essas questões produzindo *logs* sintéticos rotulados mapeados para a matriz MITRE ATT&CK, facilitando a validação de sistemas de detecção [Huang et al., 2022].

Além disso, embora a resposta a incidentes tenha tradicionalmente dependido de *playbooks* estáticos, os LLMs oferecem uma alternativa promissora para automatizar a conversão de *logs* técnicos e indicadores em planos de resposta estruturados. Esses modelos podem interpretar descrições complexas e mapeá-las para a *cyber kill chain* e sugerir ações de mitigação consistentes com frameworks estabelecidos, como NIST IR e MITRE ATT&CK [Hammar et al., 2025]. Essa mudança reflete um movimento consistente em direção à integração de ferramentas ofensivas, dados sintéticos robustos e sistemas inteligentes de suporte à decisão.

2.3. Framework MITRE CALDERA

Desenvolvido pela MITRE Corporation[1], o CALDERA é um *framework* de código aberto projetado para emulação de adversários e automação de cenários de simulação de violação e ataque, estabelecendo-se como um padrão para experimentos e validações de cibersegurança [MITRE, 2025]. Originalmente concebido para apoiar atividades de *Red Team* e operações de resposta baseadas na matriz MITRE ATT&CK, o sistema evoluiu para se tornar uma das ferramentas mais abrangentes e amplamente adotadas por profissionais.

A arquitetura da plataforma baseia-se em dois blocos de construção principais: um núcleo operacional, composto por um servidor de Comando e Controle (C2) assíncrono acessível via API REST e uma interface web, e um ecossistema de *plugins* que modulariza suas funções. Essa separação fornece a flexibilidade necessária para que a comunidade desenvolva novos comportamentos ofensivos e explore abordagens sofisticadas de tomada de decisão sem comprometer a estrutura principal.

O CALDERA gerencia a emulação de ameaças em torno de entidades conceituais centrais:

- **Agentes:** programas de *software* (por exemplo, o agente padrão *Sandcat*, escrito em GoLang) que se conectam ao servidor CALDERA em intervalos definidos para receber instruções e executar comandos;

[1] mitre.org

- **Habilidades:** ações atômicas que implementam uma técnica específica do ATT&CK. Uma habilidade define a plataforma alvo, tática, técnica, comando de execução e um comando opcional de limpeza para reverter o impacto da ação;
- **Adversários:** perfis de ameaças consistindo em habilidades encadeadas projetadas para imitar o comportamento de grupos de ameaças do mundo real;
- **Operações:** responsável pela execução de um perfil de adversário contra agentes alvos, registrando cada etapa em uma cadeia operacional.

Embora inicialmente projetado para automatizar ações de *Red Team*, o CALDERA é cada vez mais empregado para treinar *Blue Teams* e medir a eficácia das defesas existentes sob condições controladas. Um componente crítico nesse processo é o *Planner* (Planejador), que governa a lógica de progressão das habilidades e define as condições de execução necessárias, guiando efetivamente as operações ofensivas automatizadas [MITRE, 2025].

3. Trabalhos Relacionados

Esta seção revisa os principais trabalhos identificados na literatura acerca do uso de LLMs na cibersegurança, abrangendo tarefas ofensivas e defensivas, como automação de *Red Team*, planejamento MITRE ATT&CK e análise pós-incidente. Devido à vasta expansão recente da literatura sobre inteligência artificial aplicada à segurança, a seleção dos estudos aqui apresentados seguiu critérios alinhamento ao escopo do BAGLEY.

Foram priorizados trabalhos que propõem arquiteturas aplicadas, em detrimento de estudos puramente teóricos ou focados exclusivamente na detecção isolada de *malwares*. O recorte buscou ferramentas e *frameworks* que exploram ativamente a orquestração tática de ataques como emulação de adversários e CTI ou a automação de defesa como o processo de análise forense e enriquecimento de SIEM.

Como mostrado na Tabela 1, a maioria dos estudos foca em domínios isolados; além disso, embora a maioria atue como sistemas de suporte à decisão, apenas alguns [Singer et al., 2025, Sandaruwan et al., 2025] alcançam execução totalmente automatizada. Esta visão geral estabelece as fronteiras de pesquisa atuais e destaca as lacunas de integração contínua que o trabalho proposto visa preencher.

O estudo de [Liu et al., 2018] propõe o Aurora, que integra LLMs com técnicas de planejamento baseadas em PDDL. Diferentemente do foco na autonomia em tempo real, essa abordagem prioriza a previsibilidade e a reprodutibilidade, gerando sequências documentadas mapeadas diretamente para a matriz MITRE ATT&CK. Complementarmente, o *framework* SAGA [Huang et al., 2022], aborda a escassez de dados gerando *logs* sintéticos rotulados associados a campanhas de Ameaças Persistentes Avançadas (APT). O sistema combina comportamentos mapeados para o MITRE ATT&CK, fornecendo uma base relevante para validar sistemas de detecção.

O estudo de [Singer et al., 2025] explora LLMs na coordenação de ataques distribuídos em múltiplos *hosts*. Os autores propõem o Incalmo, que atua como uma camada intermediária entre o modelo e o ambiente operacional. Neste contexto, o LLM gera diretivas estratégicas de alto nível, enquanto o sistema as traduz em comandos executáveis.

No campo da resposta a incidentes, [Hammar et al., 2025] propõe soluções que combinam ajuste fino leve com Geração Aumentada por Recuperação (RAG). O objetivo

Tabela 1. Síntese comparativa dos trabalhos relacionados.

Autores	Ferramenta Base	Foco	Autonomia	Contribuição Principal
[Liu et al., 2018]	AURORA (PDDL)	Ataque (CTI)	Semi-Auto	Geração determinística de cadeia de ataque via planejamento PDDL.
[Huang et al., 2022]	SAGA (CFG)	Dados Sintéticos	Sim	Geração de *logs* de auditoria rotulados para validação de detecção de APT.
[Singer et al., 2025]	Incalmo (CALDERA)	Ofensivo	Sim	Abstração de alto nível para ataques complexos em múltiplos *hosts*.
[Hammar et al., 2025]	DeepSeek-R1	Defesa (IR)	Suporte	Planejamento de resposta com alucinação reduzida (RAG + Fine-tuning).
[Tellache et al., 2025]	RAG (Hybrid CTI)	Defesa (IR)	Suporte	Enriquecimento de alertas SIEM com CTI para planos de Resposta a Incidentes.
[Hardikar et al., 2025]	CyberSleuth	Forense	Não	Análise autônoma de rastros brutos e síntese de evidências.
[Sandaruwan et al., 2025]	Ferramenta PoC (RAG)	Vulnerabilidade	Semi-Auto	Varredura automatizada de vulnerabilidades e guias de remediação.
Este Trabalho	**BAGLEY**	**Purple Team**	**Sim**	**Orquestração autônoma de ataques multiplataforma e análise forense.**

é reduzir alucinações e gerar recomendações defensivas fundamentadas em registros de ameaças reais. Na mesma linha, [Tellache et al., 2025] utiliza RAG e CTI híbrida para enriquecer alertas de SIEM, transformando dados brutos em planos de resposta estruturados e consistentes com o *framework* NIST.

Para a análise forense, [Hardikar et al., 2025] apresenta o CyberSleuth, uma arquitetura multiagente especializada na análise de pacotes de rede e *logs* de aplicação. O sistema sintetiza evidências em linguagem natural, reduzindo o esforço manual em investigações pós-incidente. Por fim, [Sandaruwan et al., 2025] descreve uma ferramenta para varredura automatizada de vulnerabilidades em redes, utilizando LLMs para interpretar resultados de varreduras e gerar guias de remediação personalizados.

Diferente dos trabalhos supracitados, que tratam as fases ofensivas e defensivas de forma estanque, esta pesquisa as integra em um fluxo contínuo e interdependente. Ao incorporar CVEs, *Exploits* e *Payloads*, o LLM vai além das TTPs genéricas para selecionar *exploits* compatíveis com o ambiente, o que, por sua vez, melhora a análise defensiva. O projeto BAGLEY se distingue da literatura especializada por utilizar esses elementos práticos e métricas objetivas como a cobertura do MITRE ATT&CK e o tempo de resposta para fornecer uma avaliação mais realista da eficácia defensiva.

4. BAGLEY: Arquitetura do Sistema e Fluxo Operacional

Esta seção detalha o BAGLEY, uma solução integrada que operacionaliza o ciclo automatizado de *Purple Teaming* ao unificar motores cognitivos com infraestruturas de segurança consolidadas. O sistema foi projetado para superar as limitações dos testes de invasão (*penetration testing*) esporádicos, estabelecendo um ciclo de avaliação contínua que integra a emulação ofensiva do MITRE CALDERA com a análise defensiva do Wazuh SIEM[2] sob a orquestração de uma LLM.

[2]github.com/wazuh/wazuh

4.1. Arquitetura do Sistema

Conforme ilustrado na Figura 1, a arquitetura do BAGLEY estabelece a estrutura base e os protocolos de comunicação necessários para sincronizar a orquestração, a execução ofensiva e a análise defensiva. Ao organizar esses componentes em um sistema unificado, a arquitetura garante um fluxo contínuo de dados entre o núcleo cognitivo e o ambiente operacional.

O núcleo cognitivo utiliza o modelo `gemma-3-27b-it`[3] via API, operando com consistência sintática e reduzida taxa de erro na conversão de intenções táticas. Esse modelo foi selecionado por apresentar um equilíbrio entre capacidade de raciocínio, qualidade na geração de instruções e viabilidade de execução em infraestrutura local. A escolha de um modelo aberto também favorece a reprodutibilidade dos experimentos e reduz dependências de serviços proprietários, permitindo maior controle sobre o processamento.

O BAGLEY adota a Injeção de Contexto Dinâmico para unir a memória paramétrica do modelo à telemetria do Wazuh, otimizando o desempenho com respostas sob 15 segundos. Com consumo médio de 28.400 tokens de entrada e 920 de saída, a arquitetura reduz custos e complexidade em relação ao RAG convencional, garantindo uma orquestração resiliente e escalável.

Diferentemente de abordagens que exigem um treinamento extensivo como *fine-tuning*, o sistema não depende exclusivamente da memória paramétrica pré-treinada da IA; em vez disso, ele consulta dinamicamente bases de conhecimento externas para obter o contexto mais recente. Durante a fase de planejamento tático, o motor cognitivo correlaciona a intenção do operador acessando a base do *National Vulnerability Database* (NIST NVD) e o repositório contínuo de Provas de Conceito (PoCs) mantido por 0xMarcio. Essa arquitetura permite que o orquestrador valide as vulnerabilidades (CVEs) e extraia *exploits* atualizados, garantindo um alinhamento preciso com o *framework* MITRE ATT&CK sem a necessidade de re-treinar o modelo de linguagem. O *MainController* gerencia as requisições do usuário e governa as interações, enquanto o *ServerLLM* realiza

[3]huggingface.co/google/gemma-3-27b-it

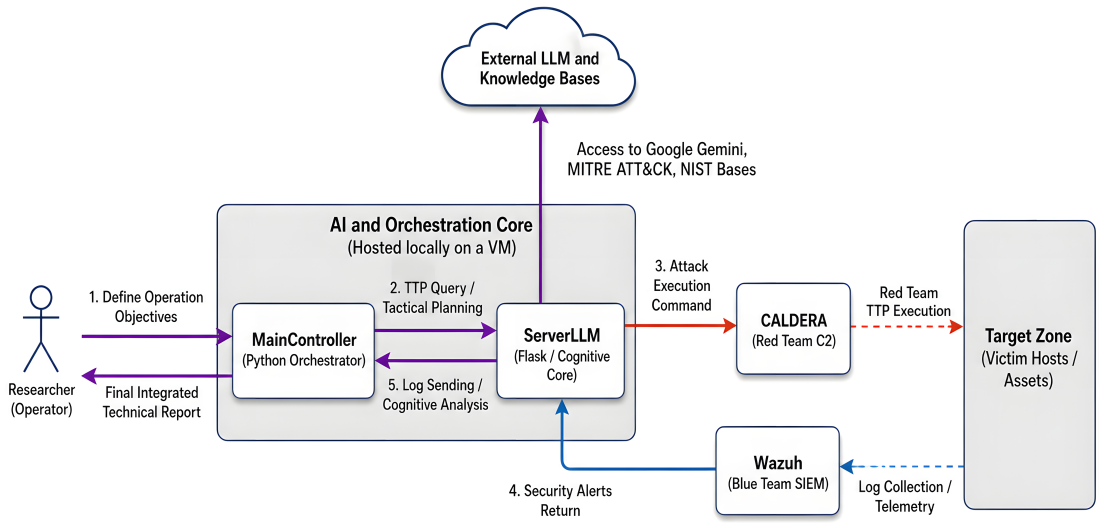


Figura 1. Arquitetura do Sistema BAGLEY.

o processamento semântico de forma autônoma.

O fluxo de trabalho é iniciado quando o operador define objetivos operacionais, que são processados pelo *MainController* e encaminhados ao *ServerLLM*. Este módulo estabelece comunicação com fontes de conhecimento externas, incluindo LLMs e bancos de dados especializados como MITRE ATT&CK e NIST, permitindo a geração de estratégias ofensivas fundamentadas em inteligência de ameaças (*threat intelligence*) atualizada.

Com base na resposta do módulo cognitivo, o *MainController* emite comandos para o *framework* CALDERA, que é responsável por executar ações ofensivas no ambiente alvo. O CALDERA atua como a infraestrutura de Comando e Controle (C2), implantando TTPs nos *hosts* vítimas dentro da zona alvo do sistema. Simultaneamente, o ambiente monitorado é supervisionado pelo Wazuh, atuando como o componente de *Blue Team*. Este sistema coleta *logs* e telemetria gerados pelos *hosts* durante a execução ofensiva, incluindo eventos de segurança relevantes.

O Wazuh foi selecionado por ser uma plataforma SIEM/XDR de código aberto amplamente utilizada em ambientes corporativos, oferecendo coleta centralizada de eventos, monitoramento de integridade de arquivos, integração com Sysmon, mecanismos flexíveis de correlação e geração de regras de detecção em formato XML. Além disso, sua API REST facilita a automação do ciclo proposto, permitindo que as regras geradas pelo sistema sejam incorporadas ao ambiente de forma padronizada.

A telemetria coletada pelo Wazuh é então enviada de volta ao núcleo de orquestração, onde o *ServerLLM* conduz uma análise cognitiva, correlacionando os eventos observados com as ações executadas. Esse processo permite a identificação de falhas de detecção, a avaliação da eficácia dos controles de segurança e a geração de recomendações de mitigação acionáveis.

Por fim, os resultados da análise são consolidados pelo *MainController* e apresentados ao operador como um relatório técnico integrado. Essa arquitetura viabiliza um ciclo contínuo de avaliação de segurança, no qual os ataques são automaticamente planejados, executados e analisados, promovendo uma integração efetiva entre as perspectivas ofensiva e defensiva.

4.2. Fluxo Operacional

O fluxo operacional do BAGLEY, conforme ilustrado no diagrama de sequência na Figura 2, destaca a interação entre o operador humano, o orquestrador, o módulo cognitivo, a infraestrutura ofensiva (*Red Team*) e o sistema de monitoramento defensivo (*Blue Team*). Este fluxo de trabalho é estruturado em três fases principais: planejamento de adversários, execução e monitoramento, e análise forense cognitiva.

4.2.1. Fase 1: Planejamento de Adversários (Red Team)

Nesta fase, o *MainController* recebe objetivos em linguagem natural e os despacha ao *ServerLLM*. O modelo Gemini analisa o contexto do alvo como sistema operacional e privilégios para selecionar a TTP mais adequada, extraindo de sua base paramétrica o comando exato e o ID MITRE. Essa etapa automatiza a inteligência de ameaças, elimi-

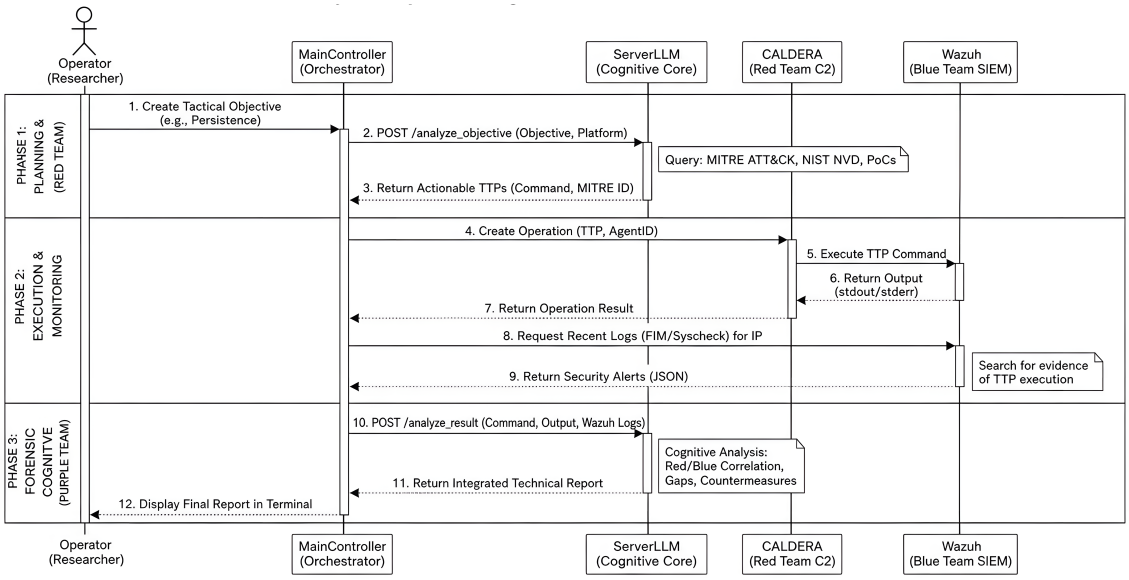


Figura 2. Diagrama de Sequência do BAGLEY.

nando a necessidade de mapeamento manual de *exploits* e garantindo que o comando gerado possua alta acurácia sintática para a plataforma específica (Bash ou PowerShell).

4.2.2. Fase 2: Execução e Monitoramento

O orquestrador consome a TTP gerada para instanciar uma operação via API do CALDERA, que atua como servidor C2. Enquanto o agente executa a manobra no *host* e retorna o `stdout`/`stderr`, o sistema realiza o *matching* dinâmico do IP do alvo com a base de agentes do Wazuh. Essa integração síncrona permite que o orquestrador capture alertas de segurança e eventos de integridade de arquivos (FIM) precisamente no intervalo de tempo da execução ofensiva, isolando a telemetria relevante.

4.2.3. Fase 3: Análise Forense Cognitiva (Purple Team)

Na consolidação, o LLM recebe o comando executado, o retorno real do terminal e os *logs* do SIEM. O sistema atua como um analista de SOC Nível 3, filtrando ruídos benignos do sistema para correlacionar evidências de detecção com a técnica emulada. O resultado é um relatório técnico que diagnostica o sucesso da invasão, aponta falhas de cobertura e fornece uma contramedida como a criação da regra XML do Wazuh, pronta para implantação, fechando o ciclo de maturidade defensiva de forma autônoma.

5. Avaliação

A avaliação concentra-se em três casos de uso principais: (i) extração *Offline* de Credenciais, que tem como alvo técnicas *Living off the Land* (LOTL) em ambientes Windows; (ii) exploração Dinâmica (PwnKit), usada para avaliar a resiliência autônoma e filtragem de ruído em alvos Linux; e (iii) reconhecimento agnóstico, que valida a abstração multiplataforma e a coleta de telemetria em ambientes mistos.

Para executar os casos de uso, os experimentos foram conduzidos em um ambiente de laboratório controlado e isolado, garantindo segurança operacional e captura de telemetria de alta fidelidade, conforme especificado na Tabela 2. Para viabilizar a reprodutibilidade experimental, os passos necessários estão integralmente documentados no repositório oficial do projeto[4]. O repositório centraliza os *scripts* de automação, os modelos de prompt do LLM e os registros (*logs*) de orquestração fundamentais para a replicação do fluxo de trabalho.

Tabela 2. Especificações Técnicas do Ambiente Experimental.

Componente	**Especificação / Versão**
Processador (vCPU)	1 Núcleo Dedicado
RAM	8 GB
HD	50 GB (Virtualizado)
Sistema Operacional	Ubuntu 24.04.3 LTS
Framework de Red Team	MITRE CALDERA v5.3.0
SIEM de Blue Team	Wazuh Manager v4.8.0 (Docker)
Motor Cognitivo	Google Gemini gemma-3-27b-it
Nós Alvos (Target Nodes)	2 VMs Heterogêneas (Win/Linux)

5.1. Caso de Uso #1: Extração Offline de Credenciais e Living off the Land

Este primeiro caso de uso simula uma tática crítica de acesso a credenciais utilizada por Ameaças Persistentes Avançadas (APTs): o roubo de bancos de dados criptográficos nativos do Windows (hives SAM e SYSTEM). O objetivo valida a capacidade do orquestrador de operar exclusivamente usando ferramentas de sistema pré-instaladas (*Living off the Land* - LOTL), evitando a injeção de binários externos.

Ao receber o objetivo em linguagem natural para extrair as hives (T1003.002), o orquestrador gerou autonomamente o comando nativo `reg save`. Conforme detalhado na Listagem 1, o modo de monitoramento X-RAY registra o ciclo de vida exato da operação, provando que o orquestrador despachou o comando e recebeu o código de saída de sucesso dentro de uma janela de 48 segundos.

Listing 1. Linha do Tempo de Execução para Extração de Hives no Windows.

```
>> Monitoring Caldera API (X-RAY Mode)...
   T+0s: Operation created (14b4fca3-eee7-49db-a4f4-868154e340a7), waiting for Planner
       ...
   T+46s: Link ID 9cc00ab0... | Status: DISCARD/WAIT
   T+48s: Link ID 9cc00ab0... | Status: SUCCESS
```

Após a execução bem sucedida, o sistema associou dinamicamente o IP alvo ao Wazuh Agent 003. Em relação à cobertura de detecção, o SIEM gerou zero alertas de segurança, caracterizando uma evasão altamente bem-sucedida pelo módulo de *Red Team*.

Para preencher essa lacuna defensiva, a incapacidade do SIEM em detectar essa extração de credenciais, o Motor Cognitivo processou os dados de execução em um tempo médio de resposta de 12 segundos. A Listagem 2 detalha a saída do motor cognitivo. O LLM faz a leitura (*parse*) precisa da saída do console para confirmar o sucesso técnico, destaca o risco de Prevenção contra Perda de Dados (DLP - *Data Loss Prevention*) das

[4]https://anonymous.4open.science/r/bagley-llm-caldera-orchestrator-4122

hives SAM/SYSTEM expostas, e gera uma regra XML do Wazuh pronta para implantação para detectar a *string* específica do PowerShell usada na evasão.

Listing 2. Relatório Forense Cognitivo para Windows.

```
INTEGRATED OPERATION REPORT | ID: 14b4fca3-eee7-49db-a4f4-868154e340a7
 Execution Diagnosis: Technical success? Yes. The stdout indicates successful completion
    . Intent: Extract SAM and SYSTEM hives.
 Exposure Analysis (DLP): SAM Hive (user hashes) and SYSTEM Hive (OS config) were
    exposed at C:\Windows\Temp\.
 Visibility & Detection: Shadow Technique detected. The absence of alerts suggests the
    attacker proactively cleared logs (T1070.001) or the SIEM lacks specific registry
    auditing.
 Countermeasure (Wazuh XML):
<group name="powershell,sam_system_hive_extraction,medium">
<rule id="100100" level="5">
<match>powershell -Command "reg␣save␣HKLM\SAM␣C:\Windows\Temp\sam.hive;␣reg␣save␣HKLM\
    SYSTEM␣C:\Windows\Temp\system.hive"</match>
<description>Detected SAM/SYSTEM Hive Extraction via PowerShell.</description>
</rule>
</group>
```

O relatório apresentado na Listagem 2 ilustra a capacidade do sistema de contextualizar *logs* brutos em inteligência de segurança estruturada através de quatro pilares analíticos:

- **Diagnóstico de Execução:** O LLM analisa a saída do console para confirmar o sucesso técnico do comando `reg save`, mapeando corretamente a intenção para acesso a credenciais via técnica T1003.002.
- **Análise de Exposição (DLP):** A análise identifica um impacto crítico na Prevenção contra Perda de Dados, observando que *hashes* de usuários sensíveis e configurações do SO foram expostos no diretório C:\windows\temp\.
- **Visibilidade e Detecção:** O sistema diagnostica uma lacuna de visibilidade, identificando que a Técnica *Shadow* (LOTL) contornou com sucesso a auditoria de registro padrão do SIEM ou os mecanismos de proteção de *logs*.
- **Contramedidas:** Para mitigar essa lacuna específica, o motor gera autonomamente uma regra XML do Wazuh projetada para acionar um alerta de nível 5 sempre que o PowerShell for usado para alvejar as *hives* SAM ou SYSTEM. A regra gerada foi aplicada ao ambiente experimental e permitiu identificar o comportamento anteriormente não detectado, demonstrando a viabilidade da abordagem proposta.

Os resultados demonstram que o BAGLEY pode efetivamente emular técnicas furtivas de LOTL e gerar autonomamente contramedidas específicas para lacunas de visibilidade que contornam as configurações padrão do SIEM. A eficácia do sistema é evidenciada por sua capacidade de transformar um cenário com 0% de cobertura de detecção inicial em um estado seguro, identificando um ponto cego crítico de detecção e gerando uma regra de mitigação.

5.2. Caso de Uso #2: Geração Dinâmica de Payload e Análise de Falsos Positivos

O segundo caso de uso eleva a complexidade ofensiva ao exigir a compilação dinâmica de código C nativo para explorar a vulnerabilidade CVE-2021-4034 (PwnKit) em um alvo Linux. O valor científico deste teste reside em avaliar a capacidade da IA de atuar como um analista de SOC Nível 2 (filtrando ruído) e em demonstrar a resiliência autônoma do orquestrador contra falhas de comunicação.

Durante a fase inicial do ataque, o agente alvo falhou em responder dentro do limite esperado, acionando um status `DISCARD/WAIT` na marca de 58 segundos, quando o servidor C2 atingiu seu limite de tempo (*timeout*). Em vez de abortar a operação, o orquestrador ativou seu mecanismo de resiliência autônoma através do *Cognitive Loop* (Loop Cognitivo). O *MainController* capturou a exceção de execução e encaminhou o contexto da falha de volta para o *ServerLLM*. O módulo cognitivo analisou a interrupção, identificando como um potencial atraso específico do ambiente ou um gargalo de compilação, e refatorou autonomamente a tarefa verificando novamente a disponibilidade do interpretador do alvo e gerando uma nova instância de operação com um ID distinto (c3e10636-...). Essa nova tentativa estratégica permitiu que o sistema contornasse o conflito de estado anterior, resultando em uma execução bem-sucedida em apenas 36 segundos, conforme capturado na Listagem 3.

Listing 3. Resiliência Autônoma e Linha do Tempo de Execução (PwnKit).

```
 Monitoring Caldera API (Attempt 1/5)...
   T+58s: Link ID c7174ebd... | Status: DISCARD/WAIT
>> Real Timeout (Agent did not respond). Retrying...
 Monitoring Caldera API (Attempt 2/5)...
   T+0s: Operation created (c3e10636-970b-46a3-bf7f-b65e5c0bda94)...
   T+36s: Link ID 720d9248... | Status: SUCCESS
```

A Listagem 4 detalha o raciocínio forense da IA. Ela isola explicitamente o ruído benigno do SO (como alertas do `snapd`) dos indicadores reais de pós-exploração, que são modificações em arquivos de configuração do serviço de impressão. Além disso, mapeia a compilação maliciosa em C diretamente para a CVE-2021-4034, fornecendo uma regra XML altamente específica para detectar futuras explorações do PwnKit.

Listing 4. Relatório Forense Cognitivo com Filtragem de Ruído (CVE-2021-4034).

```
INTEGRATED OPERATION REPORT | ID: c3e10636-970b-46a3-bf7f-b65e5c0bda94
 Execution Diagnosis: True. Intent: Gain elevated privileges via pkexec (TA0004).
 Exposure Analysis (DLP): Potential root-level compromise. Attacker demonstrated
    capability to compile and execute arbitrary C code.
 Visibility & Detection: FILTER NOISE. Wazuh logs for snapd are noise. RELEVANCE: /etc/
    cups/subscriptions.conf modification is a strong indicator of malicious activity.
 MITRE & CVE Correlation:
- MITRE Technique: T1548.002 - Abuse Elevation Control Mechanism: pkexec.
- Vulnerability: CVE-2021-4034 (Polkit pkexec).
 Countermeasures (Wazuh XML):
<group name="pkexec_abuse,sid:1000001,detect:true">
  <rule id="1000001" level="11">
    <match>COMMAND:gcc -fPIC -shared -o /tmp/evil.so /tmp/evil.c</match>
    <description>Detected attempt to exploit pkexec vulnerability (CVE-2021-4034).</
        description>
  </rule>
</group>
```

O relatório detalhado na Listagem 4 destaca a capacidade do sistema de isolar ameaças críticas durante o evento de escalonamento de privilégios via PwnKit:

- **Diagnóstico de Execução:** O LLM confirma o sucesso técnico da operação, identificando corretamente a intenção de obter acesso *root* através do utilitário pkexec, associado à tática TA0004.
- **Análise de Exposição (DLP):** A análise identifica um risco crítico de comprometimento em nível *root*, já que o atacante demonstrou com sucesso a capacidade de compilar e executar código C arbitrário no alvo.

- **Visibilidade e Detecção:** Atuando como um analista de SOC avançado, o sistema filtra *logs* de fundo benignos do snapd e identifica a modificação de /etc/cups/-subscriptions.conf como o indicador de alta relevância de atividade maliciosa.
- **Correlação MITRE e CVE:** O relatório realiza um mapeamento técnico preciso, vinculando o ataque diretamente à vulnerabilidade PwnKit (CVE-2021-4034) e à técnica MITRE T1548.002.
- **Contramedidas:** O motor gera autonomamente uma regra XML de nível 11 do Wazuh adaptada especificamente para detectar a string de comando GCC usada para a compilação da biblioteca maliciosa, fechando a lacuna de visibilidade identificada.

Este cenário validou a capacidade do motor cognitivo de operar como um analista avançado de SOC. A eficácia do sistema é evidenciada por sua alta precisão analítica e resiliência operacional, mantendo a continuidade dos testes através de uma recuperação autônoma de um *timeout* de execução de 58 segundos. Ao filtrar com precisão ruídos benignos do sistema, como alertas do `snapd`, consegue isolar indicadores relevantes de pós-exploração.

5.3. Caso de Uso #3: Orquestração Multiplataforma e Abstração de SO

Este caso de uso demonstra a abstração agnóstica de sistema operacional da arquitetura e valida as métricas objetivas propostas, especificamente, a cobertura de detecção e o tempo de resposta. Um único objetivo tático foi submetido em linguagem natural: simular uma fase de reconhecimento pós-comprometimento e preparação (*staging*) para exfiltração de dados (T1082, T1074). O sistema foi instruído a coletar tabelas de roteamento, conexões de rede e processos em execução, ocultando a telemetria coletada em um arquivo de registro temporário e oculto no diretório do sistema.

Como mostrado na Listagem 5, o processo começa carregando o motor cognitivo, mapeando os alvos heterogêneos disponíveis (Linux e Windows) e recebendo o objetivo tático diretamente do operador.

Listing 5. Inicialização de Rede e Identificação de Alvos Mistos.

```
>> SCANNING NETWORK...
NUM  PAW      GROUP   PLATFORM      IP
1    gymlvu   red     linux    192.168.1.65
2    mnwoia   red     windows  192.168.56.102
>> Select IDs (e.g., 1,2), Group (e.g., red) or 'all': red
```

O orquestrador LLM interpretou com sucesso a intenção e gerou *payloads* nativos distintos para o ambiente heterogêneo em tempo real. No alvo Linux, o objetivo foi automaticamente traduzido em comandos Bash (`ip route`, `netstat`). Em relação à cobertura de detecção dentro deste cenário controlado, o SIEM mapeou com sucesso todos os artefatos emulados, gerando 5 alertas de Monitoramento de Integridade de Arquivos. A Listagem 6 demonstra a análise forense do Motor Cognitivo sobre esta execução no Linux.

Listing 6. Relatório Forense Cognitivo para o Alvo Linux.

```
INTEGRATED OPERATION REPORT | ID: 836d58ad-dd96-48be-ac1f-33381d5eaa22
>> Execution Diagnosis: True. Intent: gather system information (TA0007: Discovery -
    T1082, T1016).
```

```
>> Exposure Analysis (DLP): Routing Table, Network Connections and Process List exposed
    via /tmp/.sys_telemetry.log.
>> Visibility & Detection: FILTER NOISE: logs related to snap packages are noise.
    HIGHLIGHT RELEVANCE: [FIM] Modified: /etc/cups/subscriptions.conf.
>> Countermeasures (Wazuh XML):
<group name="command_detection,">
  <rule id="100100" level="7">
    <match>ps aux > /tmp/.sys_telemetry.log</match>
    <description>Detection of reconnaissance and data exfiltration.</description>
  </rule>
</group>
```

Simultaneamente, o orquestrador adaptou a mesma intenção em *cmdlets* modernos do PowerShell para o alvo Windows, concluindo a execução em 42 segundos. Apesar do sucesso técnico no Windows, o SIEM gerou apenas 1 alerta. Para lidar com essa lacuna crítica de detecção, o Motor Cognitivo processou ambos os fluxos de dados em um tempo médio de resposta de 14 segundos para gerar autonomamente mitigações acionáveis. A Listagem 7 exibe o relatório forense resultante.

Listing 7. Relatório Forense Cognitivo para Windows (Evasão via LOLBin).

```
INTEGRATED OPERATION REPORT | ID: 9aa2982b-67c5-491b-a5ce-35ae4ed24403
 Execution Diagnosis: True. Command intent: gather system telemetry data and append it
    to a log file (TA0043: Reconnaissance).
 Exposure Analysis (DLP): Exposed network topology and communication patterns.
 Visibility & Detection: The attack evaded detection due to the use of a Living Off The
    Land Binary (LOLBin) - PowerShell.
 Countermeasures (Wazuh XML):
<group name="powershell,reconnaissance,">
  <rule id="100100" level="7">
    <match>Get-NetRoute | Format-List | Out-File</match>
    <description>PowerShell Telemetry Collection - Reconnaissance Attempt</description>
  </rule>
</group>
```

Os resultados deste estudo de caso consolidam a orquestração como o principal diferencial técnico do sistema BAGLEY. A eficácia da arquitetura é validada por sua capacidade de expor posturas de segurança díspares em ambientes heterogêneos através de métricas objetivas. Enquanto o alvo Linux forneceu evidências abrangentes de detecção através de múltiplos alertas de segurança, o ambiente Windows revelou uma lacuna significativa de visibilidade para a mesma intenção tática.

6. Considerações Finais

Para combater a velocidade das ameaças modernas, o BAGLEY preenche a lacuna entre a defesa manual e a resposta autônoma ao orquestrar LLMs dentro dos ecossistemas do CALDERA e do Wazuh. Este estudo demonstra que a IA devidamente governada viabiliza a validação de segurança contínua, em vez de avaliações pontuais.

Como trabalhos futuros, a arquitetura será expandida para um modelo agnóstico de integração, permitindo o acoplamento simplificado com outros sistemas de monitoramento e EDR além do Wazuh, por meio de conectores padronizados. Outro ponto interessante vem também da implementação de geração autônoma de fluxogramas táticos nos relatórios finais, fornecendo uma visualização clara das etapas do ataque e dos caminhos de exploração percorridos.

Referências

Nguyen, T. T. *et al.* (2021). Deep reinforcement learning for cyber security. *IEEE Transactions on Neural Networks and Learning Systems*.

Jaffal, N. O., Alkhanafseh, M., and Mohaisen, D. (2025). Large Language Models in cybersecurity: A survey of applications, vulnerabilities, and defense techniques. *AI*, 6(9):216.

Singer, J. *et al.* (2025). On the feasibility of using LLMs to autonomously execute multi-host network attacks. *arXiv preprint 2501.16466*.

Liu, Y. *et al.* (2018). AURORA: Automated Adversary Emulation via PDDL and Large Language Models. In *Proc. 24th ACM SIGKDD*.

Hammar, K. *et al.* (2025). Incident Response Planning Using a Lightweight Large Language Model with Reduced Hallucination. *arXiv preprint arXiv:2508.05188*.

Tellache, A. *et al.* (2025). Advancing Autonomous Incident Response: Leveraging LLMs and Cyber Threat Intelligence. *arXiv preprint arXiv:2508.10677*.

Hardikar, S. *et al.* (2025). CyberSleuth: A Multi-Agent Architecture for Autonomous Forensic Analysis. In *Proc. Int. Conf. on Information Security and Cryptology (ICISC)*.

Huang, L. *et al.* (2022). SAGA: Synthetic Audit Log Generation for APT Campaigns. In *Proc. ACM SIGSAC CCS*.

Sandaruwan, M. T., Wijayanayake, J., and Senanayake, J. (2025). Integrating Large Language Models for automated vulnerability scanning and reporting in network hosts. In *IEEE SCSE*.

MITRE (2025). CALDERA Documentation. Online. Available: https://caldera.readthedocs.io/

Miller, D. *et al.* (2018). Automated Adversary Emulation: A Case for Planning and Acting with Unknowns.

Yin, Z. *et al.* (2025). Digital Forensics in the Age of Large Language Models. *arXiv preprint 2504.02963*.

Wang, L. *et al.* (2024). From Sands to Mansions: Towards Automated Cyberattack Emulation with Classical Planning and Large Language Models. *arXiv preprint 2407.16928*.

Carlini, N. *et al.* (2026). Assessing Claude Mythos Preview's cybersecurity capabilities. Anthropic. Online. Disponível em: `https://red.anthropic.com/2026/mythos-preview/`.