Uma Abordagem Baseada em LLMs Open-Weight para Detecção e Classificação de Vulnerabilidades em Relatórios OSINT

Daniel R. Barros[1], João P. Lima[1], Carlos Caminha[1], José M. Monteiro[1], Javam Machado[1].

[1]Laboratório de Sistemas e Bancos de Dados (LSBD)
Departamento de Computação (DC)
Universidade Federal do Ceará (UFC) – Fortaleza, CE – Brasil.

{daniel.rezende,joaopaulo.lima}@lsbd.ufc.br

{carlos.caminha,jose.monteiro,javam.machado}@lsbd.ufc.br

Resumo. *A pesquisa proposta tem como objetivo avaliar Modelos de Linguagem de Grande Porte (LLMs) na tarefa de análise de resultados obtidos a partir de testes de segurança em ambientes online. O foco central está relacionado à interpretação automatizada de relatórios de exploração gerados por ferramentas OSINT, utilizados para identificar vulnerabilidades e, assim, reduzir a dependência de análises manuais exaustivas. O estudo aborda a lacuna associada à complexidade da inspeção desses relatórios, cuja natureza altamente técnica dificulta a identificação, a classificação e a priorização de falhas de segurança. A proposta também é motivada por diretrizes recentes da OWASP Foundation, que destacam a necessidade do uso de LLMs confiáveis e seguros em aplicações críticas. Para a realização da classificação de vulnerabilidades nos ambientes online analisados, é proposta uma taxonomia de integração unificada composta por cinco categorias Macro, construída a partir da integração de metodologias consolidadas dos frameworks OWASP Top 10 e FIRST CVSS. Os resultados demonstraram que modelos como qwen3.6-35b-a3b e gemma-4-26b-a4b apresentaram os melhores desempenhos gerais na avaliação, destacando-se nas métricas de F1-Score, Recall e Precision. Além disso, modelos compactos como o gemma-4-e2b também obtiveram resultados competitivos, indicando que LLMs menores podem representar alternativas viáveis em cenários com restrições computacionais.*

1. Introdução

O panorama contemporâneo da Internet é composto majoritariamente por aplicações web complexas, altamente interativas e dependentes da comunicação contínua entre cliente e servidor. Esses serviços são responsáveis por suportar funcionalidades críticas como autenticação de usuários, transações financeiras, processamento de dados pessoais e compartilhamento de informações públicas e privadas. Nesse contexto, o conteúdo apresentado é frequentemente gerado dinamicamente e personalizado em tempo real para cada usuário, fazendo com que grande parte das informações processadas possuam caráter sensível ou confidencial.

Como destacado pelo relatório quantitativo [World Economic Forum 2025], o número de ataques direcionados a aplicações *online* e infraestruturas sensíveis aumentaram significativamente nas últimas décadas, tanto em frequência quanto em sofisticação. Paralelamente, a crescente dependência global por ambientes virtualizados e esferas tecnológicas fazem com que grande parte das atividades econômicas, sociais e governamentais estejam diretamente associadas à disponibilidade e segurança desses sistemas. Nesse cenário, incidentes cibernéticos podem resultar em impactos de larga escala, comprometendo serviços essenciais, informações sensíveis e operações estratégicas. Dessa forma, torna-se evidente que abordagens puramente reativas são insuficientes, sendo necessário o desenvolvimento contínuo de técnicas, metodologias e ferramentas especializadas capazes de prevenir, detectar e corrigir essas ameaças.

Como forma de mitigação contra essas ameaças cibernéticas, podemos destacar a estratégia da análise de superfícies de ataque, uma prática fundamental adotada por muitos profissionais de Segurança Cibernética, cujo objetivo consiste em mapear e avaliar os pontos de exposição acessíveis externamente em sistemas computacionais [Theisen et al. 2018]. Para isso, uma das abordagens adotadas por especialistas em cibersegurança consiste na realização de explorações OSINT (*Open Source Intelligence*), utilizando ferramentas especializadas para coletar informações técnicas específicas sobre serviços expostos, versões de *software*, configurações inadequadas e possíveis vulnerabilidades presentes nos ambientes analisados. Entretanto, uma das principais complicações dessa abordagem reside no fato de que essas ferramentas frequentemente produzem resultados excessivamente volumosos, o que pode gerar interpretações ambíguas e dificultar a correção de vulnerabilidades existentes [Roy et al. 2022].

Nesse contexto, este trabalho propõe a utilização de Modelos de Linguagem de Grande Porte (*Large Language Models* - LLMs) como uma solução de apoio à identificação, categorização e priorização de vulnerabilidades em ambientes *online*. Por meio da seleção de alguns dos modelos mais relevantes da atualidade, o estudo busca identificar a tecnologia LLM mais adequada para reduzir o esforço manual e aumentar a consistência e a confiabilidade das análises realizadas. Entre as principais contribuições desta pesquisa, destacam-se: (i) a construção de um *dataset* de vulnerabilidades baseado em explorações OSINT conduzidas em sites e aplicações web reais; (ii) o desenvolvimento de uma taxonomia de integração entre OWASP e CVSS para interpretação de evidências OSINT; (iii) a realização de um *Benchmark* para avaliação de LLMs *Open-Weight*; (iv) a definição de um *framework* de validação fundamentado na abordagem *LLM-as-a-Judge*; e (v) a realização de uma análise comparativa entre especialistas humanos e os modelos de IA selecionados.

2. Fundamentação Teórica

Desde 2023, os LLMs como o GPT, passaram a ocupar uma posição central de destaque nos avanços da Inteligência Artificial. Esses modelos representam o acumulo de décadas de progresso nas áreas de Aprendizado Profundo (*Deep Learning*) e Processamento de Linguagem Natural (PLN). Este capítulo tem o objetivo de apresentar algumas definições breves sobre conceitos básicos dessas tecnologias, para possibilitar uma compreensão mais abrangente sobre seu funcionamento.

Tokens: São as unidades básicas de texto processadas pelos modelos de linguagem, podendo representar palavras, subpalavras ou símbolos. A tokenização consiste na conversão do texto em representações numéricas, adequadas ao processamento computacional. O processo de tokenização é exemplificado em [Hugging Face 2023][1].

Embeddings: São representações vetoriais que capturam as relações semânticas entre palavras em um espaço multidimensional, de modo que termos semanticamente semelhantes tendem a apresentar vetores próximos.

Self-Attention: O mecanismo de autoatenção, permite ao modelo atribuir diferentes pesos aos elementos de uma sequência ao modelar dependências contextuais. Essa abordagem substituiu, em grande escala, as arquiteturas recorrentes tradicionais.

Context Window: Refere-se à quantidade máxima de *tokens* que o modelo pode considerar simultaneamente durante o processo de inferência.

Conforme descrito por [Vaswani et al. 2017], o processo de geração de texto em LLMs funciona de forma autoregressiva, estimando iterativamente a probabilidade dos próximos *tokens* com base no contexto previamente observado. A cada predição, o *token* selecionado é incorporado à sequência de entrada, repetindo esse processo até que um critério de parada seja atingido, como um *token* de término ou o comprimento máximo da sequência [Kaplan et al. 2020].

Apesar do grande potencial da tecnologia dos LLMs, os modelos amplamente utilizados na indústria ainda apresentam limitações críticas, conforme detalhado em [OWASP LLMs Risk 2025]. Essas limitações impactam diretamente a confiabilidade de sua aplicação em domínios sensíveis, como o caso da área da Segurança Cibernética, nos quais a precisão e robustez são requisitos fundamentais. Dentre os principais riscos, destacam-se:

1. **`Inconsistência nas Respostas:`** modelos de uso geral podem produzir respostas divergentes para consultas semanticamente equivalentes, prejudicando a estabilidade e a confiabilidade analítica.
2. **`Persistência no Erro ("Teimosia"):`** em certos casos, modelos generalistas insistem em respostas incorretas, mesmo quando confrontados com correções, podendo gerar falsos positivos e análises de alto impacto.
3. **`Limitação da Janela de Contexto:`** a quantidade de informações que o modelo pode processar simultaneamente é finita, o que resulta em respostas incompletas ou inviáveis quando tarefas envolvem grande volume de dados.
4. **`Informação Falsa e Alucinações:`** modelos generalistas frequentemente apresentam respostas factualmente incorretas, especialmente em domínios técnicos que exigem precisão rigorosa.

A ***OWASP Top 10 for LLMs***[2] destaca riscos como validação inadequada de saídas (**LLM05**), geração de conteúdo incorreto (**LLM09**), limitações semânticas (**LLM08**), consumo excessivo de recursos (**LLM10**), raciocínio falho (**LLM06**), comprometimento dos dados de treinamento (**LLM04**) e exposição de informações sensíveis (**LLM02**).

[1] https://huggingface.co/spaces/Xenova/the-tokenizer-playground
[2] https://genai.owasp.org/llm-top-10/

3. Trabalhos Relacionados

O avanço dos Grandes Modelos de Linguagem impulsionam o desenvolvimento de novas abordagens aplicadas à cibersegurança, especialmente em tarefas relacionadas à detecção de vulnerabilidades, análise de ameaças e interpretação automatizada de informações técnicas. Diversos estudos recentes investigam o uso desses modelos em cenários como análise de código-fonte, classificação de vulnerabilidades e extração de técnicas de ataque. Esta seção apresenta alguns trabalhos relacionados à utilização de LLMs em tarefas de segurança ofensiva e análise de vulnerabilidades, destacando suas abordagens, características e contribuições em relação à proposta desta pesquisa.

O estudo proposto por [Li et al. 2026] avalia cinco métodos atuais especializados baseados em modelos LLM e dois analisadores estáticos tradicionais, usando um *Benchmark* interno com 222 vulnerabilidades reais e 24 projetos *open-source*. Os resultados mostram baixa taxa de *Recall* para métodos baseados em LLM, porém eles detectam mais vulnerabilidades únicas do que ferramentas tradicionais, embora ambos apresentem altas taxas de falsos positivos. Além disso, os autores afirmam que os métodos baseados em LLM são computacionalmente caros, com custos que variam de centenas de milhares a centenas de milhões de *tokens* e tempos de execução de horas a dias, destacando limitações de robustez, confiabilidade e escalabilidade.

A ferramenta desenvolvida por [Chopra et al. 2026], denominada ChatNVD, utiliza recursos de suporte baseados em Grandes Modelos de Linguagem (LLMs), através da utilização do *National Vulnerability Database* (NVD) para melhorar o acesso às informações de vulnerabilidades. Variantes da ferramenta foram desenvolvidas com GPT-4o Mini, LLaMA 3 e Gemini 1.5 Pro. Sua avaliação é baseada em um *Benchmark* de consultas estruturadas a partir de registros reais de CVEs, abrangendo atributos temporais, descritivos e métricos. GPT-4o Mini demonstrou desempenho superior, com precisão de mais de 92% e baixa taxa de alucinações e erros. O estudo destaca o potencial dos *workflows* com LLMs leves e baseados em recuperação para apoiar a gestão e decisões operacionais em segurança cibernética.

A pesquisa conduzida em [Cuong Nguyen et al. 2025] avalia métodos de extração de *Cyber Threat Intelligence* (CTI) para identificar técnicas de ataque em relatórios, utilizando o *framework MITRE ATT&CK* e modelos LLMs, como Llama2. Durante os estudos, foram identificados desafios relevantes como desequilíbrio de classes, *overfitting* e complexidade específica do domínio, que prejudicam a extração precisa das técnicas. Porém, a proposta de um pipeline em duas etapas através da sumarização dos relatórios por LLM e processamento com SciBERT re-treinado em *dataset* reequilibrado e aumentado, resultou em melhorias expressivas nas métricas F1, alcançando acima de 90% para várias técnicas.

Tabela 1. Comparação entre Trabalhos Relacionados

Trabalho	Objetivo	Dados	Abordagem
[Li et al. 2026]	Detecção de Falhas	Projetos Open-Source	LLMs e Análise Estática
[Chopra et al. 2026]	Interpretação de CVEs	Base NVD	LLMs com recuperação
[Cuong Nguyen et al. 2025]	Extração CTI	Relatórios CTI	LLMs + SciBERT
Estudo Atual	Explorações OSINT	Relatórios Técnicos Reais	LLMs Open-Weight

Embora os estudos analisados demonstrem avanços relevantes na aplicação dos Grandes Modelos de Linguagem em tarefas relacionadas à cibersegurança, observa-se que a maior parte das abordagens concentra-se em bases antigas, análise de código-fonte ou extração de informações em relatórios CTI. Dessa forma, ainda existe uma lacuna quanto à utilização de LLMs especializados para interpretação automatizada de relatórios OSINT, produzidos durante processos de exploração de superfícies de ataque sobre alvos reais. Nesse contexto, a proposta deste estudo busca contribuir para essa área através da avaliação de modelos *Open-Weight* nesse contexto, além da nova taxonomia unificada de vulnerabilidades e avaliação profissional especializada em relatórios técnicos legítimos.

4. Metodologia

Este trabalho caracteriza-se como uma pesquisa experimental aplicada, estruturada em três etapas principais. A primeira consistiu na construção de um *dataset* próprio, desenvolvido a partir de dados reais obtidos por meio de explorações controladas em ambientes *online* vinculados a uma entidade tecnológica federal, garantindo maior confiabilidade e relevância ao conjunto de dados. A segunda etapa concentrou-se na modelagem do conhecimento técnico, integrando metodologias consolidadas como a [OWASP Foundation 2025] e o CVSS 3.1 [FIRST 2019], resultando na definição de cinco categorias Macro de vulnerabilidades para estruturar e interpretar os relatórios técnicos. Por fim, a terceira etapa abordou a avaliação dos Grandes Modelos de Linguagem para avaliar a aptidão na tarefa de análise automatizada de vulnerabilidades, buscando otimizar a interpretação de relatórios de cibersegurança e produzir análises mais precisas, consistentes e alinhadas às demandas do domínio *online*.

4.1. A Construção dos Dados

A etapa inicial da pesquisa consistiu na exploração em larga escala de centenas de alvos reais, compostos por sites e aplicações web pertencentes a um parque tecnológico universitário. Os administradores dos ambientes firmaram acordo com a equipe de pesquisa e autorizaram formalmente a realização dos experimentos.

As explorações dos ambientes foram realizadas por meio da estratégia Open Source Intelligence (OSINT), que se refere à obtenção e análise de informações provenientes de fontes abertas, acessíveis legalmente e sem restrições de segurança. Conforme [Carvalho 2023], essa abordagem baseia-se na coleta não clandestina de dados a partir de diferentes fontes públicas. No contexto da exploração de alvos na Internet, como sites e aplicações web, essa prática caracteriza-se como OSINT passiva, uma vez que não houve interação direta com o alvo, utilizando apenas informações previamente disponíveis. Nesse caso, a URL do domínio do site constituiu a única informação necessária utilizada pela ferramenta de exploração durante o processo de reconhecimento.

Ao todo, foram conduzidas explorações em 344 alvos reais, resultando na construção de um *dataset* composto por relatórios técnicos gerados a partir de execuções controladas da ferramenta Web xKaliBurr [Barros et al. 2024], utilizada como mecanismo de análise de superfícies de ataque em aplicações *online*. Um *Toy Example* [3] de relatório técnico "bruto", gerado pela ferramenta exploratória encontra-se disponível no repositório Git do projeto de 2024.

[3] `github.com/xKaliBurr/SBSeg2024/blob/main/report_Juice_Shop.txt`

Cada um dos 344 alvos analisados resultou em seu respectivo relatório técnico bruto, com cada relatório contendo aproximadamente entre 400 e 800 linhas de informações técnicas relacionadas às características dos ambientes avaliados, totalizando cerca de 206 mil linhas de dados. Esses dados representam fontes que descrevem potenciais vulnerabilidades e falhas de configuração identificadas durante as explorações, o volume expressivo de informações evidencia a complexidade do processo de interpretação manual e reforça a necessidade de abordagens automatizadas para apoio à análise de cibersegurança.

A partir desses dados, foi realizada uma etapa de seleção para coletar exemplos representativos de vulnerabilidades recorrentes presentes no domínio *online* avaliado. Foram analisados relatórios de exploração de ambientes com elevado grau de comprometimento estrutural, caracterizado pela presença de *software*s desatualizados, exposição de subdiretórios sensíveis e indícios de comprometimento ativo. Dessa forma, foi possível identificar as principais causas de explorações e ataques a esses ambientes, permitindo a rotulação das falhas e a proposição de correções para as vulnerabilidades identificadas.

Ao final do processo, foram rotuladas 257 amostras de vulnerabilidades, distribuídas entre vinte relatórios técnicos "interpretados", selecionados para compor o conjunto de *Ground Truth* para avaliação futura dos modelos LLMs *Open-Weight*. Os relatórios interpretados possuem elevado valor técnico, por representarem atividades centrais de triagem de vulnerabilidades e resposta a incidentes em Segurança Cibernética. Todos os relatórios interpretados foram construídos de forma padronizada, com cada amostra de vulnerabilidade descrita pelos campos a seguir:

- **ID**: Identificador sequencial responsável por demarcar cada achado presente no relatório ténico bruto.
- **Alvo**: URL ou endereço IP do ambiente avaliado.
- **Título**: Descrição sucinta da vulnerabilidade identificada.
- **Descrição**: Explicação técnica detalhada da falha e de sua relevância.
- **Evidência**: Referência direta ao trecho do relatório bruto onde a vulnerabilidade foi identificada.
- **Impacto**: Análise dos impactos nos pilares de Confidencialidade, Integridade e Disponibilidade.
- **Severidade**: Classificação do achado com base no CVSS Score.
- **Recomendação**: Orientações para mitigação ou remediação da vulnerabilidade.
- **Vulnerabilidade_Macro**: Classificação proposta pelo projeto para correlacionar OWASP Top 10 e CVSS (detalhadas na seção seguinte).
- **OWASP_Top_10**: Categoria OWASP Top 10:2025 associada ao achado.
- **CVSS_Score**: Pontuação de severidade no intervalo de 0.0 a 10.0.
- **CVSS_Vector**: Vetor completo de exploração segundo a metodologia CVSS v3.0.

A construção e a rotulação dos dados deste conjunto foram conduzidas sob a supervisão do analista pleno em Segurança Cibernética com sólida experiência prática em identificação, classificação, priorização e gerenciamento de vulnerabilidades em ambientes reais de produção. O processo de rotulação manual foi realizado com base em referências técnicas consolidadas e amplamente reconhecidas pela comunidade de cibersegurança, assegurando elevado rigor metodológico, consistência analítica e alinhamento com práticas profissionais adotadas em avaliações de segurança contemporâneas.

4.2. As Cinco Categorias Macro

Embora a literatura disponha de taxonomias consolidadas para classificação de vulnerabilidades, como o GitHub Security Advisories (GHSA) [GitHub Security Lab 2025], o Common Vulnerabilities and Exposures (CVE) e o Common Weakness Enumeration (CWE) [MITRE Corporation 2025], essas abordagens não estabelecem uma estrutura intermediária para organizar as evidências produzidas durante as explorações OSINT e relacioná-las simultaneamente à classificação e à priorização das falhas. Para suprir essa lacuna, este trabalho propõe cinco categorias Macro de caráter operacional, com o objetivo de estruturar as evidências coletadas, facilitar sua interpretação e correlacionar os apontamentos.

A modelagem foi baseada nas quatro primeiras categorias da [OWASP Foundation 2025], amplamente reconhecidas pela relevância das vulnerabilidades em aplicações web, complementadas pelas métricas de severidade do CVSS v3.1, mantidas pela [FIRST 2019]. Dessa forma, as categorias Macro integram aspectos qualitativos da OWASP e quantitativos do CVSS, constituindo uma camada intermediária para organização e interpretação das evidências extraídas dos relatórios OSINT. Por possuírem finalidade operacional, essas categorias podem apresentar relações de especialização entre si, refletindo a natureza das evidências observadas durante a exploração da superfície de ataque. As cinco categorias Macro são definidas a seguir:

`MACRO 01 – Information Disclosure:` Refere-se à exposição indevida de informações sensíveis relacionadas a componentes internos de sistemas, permitindo que atacantes obtenham dados relevantes para atividades de reconhecimento e exploração do alvo. Conforme discutido por [Theisen et al. 2018], a superfície de ataque pode ser ampliada quando aplicações ou serviços divulgam informações que deveriam permanecer restritas, favorecendo a identificação de vetores de intrusão. Essa falha também viola o *Need-to-Know Principle* (Princípio da Necessidade de Conhecer) [Bell and LaPadula 1973]. Além disso, essa categoria está associada à OWASP Top 10:2025 – A02: *Security Misconfiguration*, uma vez que normalmente decorre da exposição de informações como tipo e versão do servidor web, linguagem de programação e *framework* utilizado, versões de bibliotecas e sistemas operacionais, detalhes internos da arquitetura de rede, caminhos de diretórios, variáveis de ambiente e parâmetros de depuração.

`MACRO 02 – Directory Traversal:` Descreve falhas que permitem acesso não autorizado a arquivos sensíveis, diretórios ocultos ou recursos internos do servidor por meio da manipulação inadequada de caminhos e controles de acesso. Essa condição viola o Modelo de Controle de Acesso Discricionário (DAC) proposto por [Lampson 1974], ao possibilitar a obtenção de recursos restritos via manipulação de parâmetros de entrada. A categoria relaciona-se à OWASP Top 10:2025 – A01: *Broken Access Control* e pode ser identificada pela análise de requisições web e códigos HTTP definidos pela RFC 7231 [Fielding and Reschke 2014], incluindo respostas HTTP 200 para diretórios sensíveis, mensagens de erro com caminhos absolutos, exposição de arquivos de configuração, *directory listing* e acesso indevido a arquivos de *log*, *backups* ou *scripts* administrativos.

MACRO 03 – Outdated *softwares*: Expõe o uso de componentes desatualizados contendo vulnerabilidades conhecidas e sem correções aplicadas, comprometendo o *Defense in Depth* (Princípio de Defesa em Profundidade) [National Security Agency 2012], uma vez que falhas em um único componente podem impactar toda a arquitetura da aplicação. Essa categoria relaciona-se à OWASP Top 10:2025 – A03: *software Supply Chain Failures* e pode ser identificada por técnicas de *fingerprinting*, incluindo inspeção de cabeçalhos HTTP, análise temporal de campos como *Last-Modified*, correlação de versões com bases públicas de vulnerabilidades e detecção de padrões de resposta HTTP associados a tecnologias obsoletas.

MACRO 04 – Infrastructure Disclosure: Refere-se à exposição indevida de informações sobre a infraestrutura tecnológica de uma organização, como arquitetura de rede, serviços internos, subdomínios e convenções de nomenclatura, permitindo a elaboração de ataques direcionados. Essa condição viola o *Need-to-Know Principle* (Princípio da Necessidade de Conhecer) [Bell and LaPadula 1973] e relaciona-se à OWASP Top 10:2025 – A02: *Security Misconfiguration*, normalmente decorrente de configurações inadequadas de exposição. Sua identificação ocorre por técnicas de coleta e correlação de informações públicas, incluindo análise de registros DNS reversos, enumeração de subdomínios, mapeamento de topologia de rede e detecção de convenções de nomenclatura associadas a ambientes internos, como "*dev*", "*admin*", "*intranet*" e "*backup*".

MACRO 05 – Weak SSL/TLS Configuration: Por fim, esta classe descreve falhas na implementação ou configurações inadequadas de mecanismos criptográficos responsáveis pela proteção da comunicação cliente-servidor, incluindo o uso de protocolos obsoletos, algoritmos fracos e certificados digitais inválidos. Essas vulnerabilidades comprometem os princípios da Tríade CID [National Institute of Standards and Technology 1995] e podem favorecer ataques *Man-in-the-Middle* (MitM) [Anderson 2010]. A categoria relaciona-se à OWASP Top 10:2025 – A04: *Cryptographic Failures* e pode ser identificada pela análise de serviços SSL/TLS expostos, envolvendo verificação de versões de protocolos, *cipher suites*, validade de certificados digitais e presença do cabeçalho *Strict-Transport-Security (HSTS)*.

A Figura 1 apresenta o fluxo de análise das superfícies de ataque realizado pela ferramenta Web xKaliBurr, destacando os dados obtidos por cada módulo de exploração. A partir desse mapeamento, é possível relacionar os resultados produzidos por cada módulo às categorias Macro de vulnerabilidades correspondentes, identificando quais classes recebem informações provenientes de cada ferramenta utilizada na avaliação de segurança.

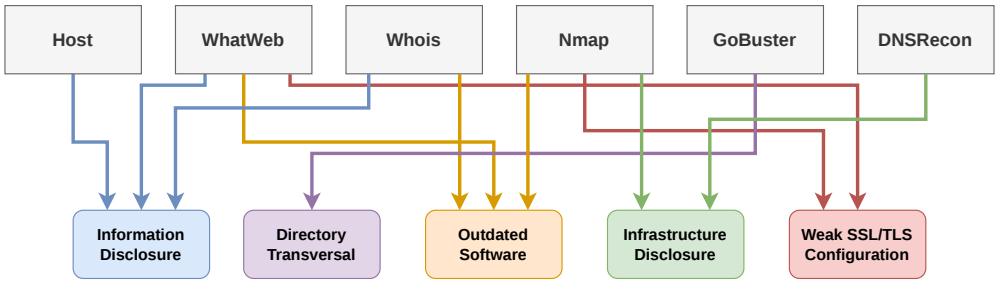

Figura 1. Coleta de Informações por Categoria Macro via OSINT

4.3. O Processo de Seleção de Modelos

Para compor o *Benchmark*, selecionamos quinze LLMs *Open-Weight* bem posicionados no ranking da plataforma *LMArena*[4] (considerando a classificação disponível em 01/05/2026), abrangendo modelos de diferentes famílias e escalas de parâmetros, desde candidatos compactos até arquiteturas de alto desempenho. Todos utilizaram quantização Q4_K_M, uma técnica amplamente empregada para reduzir o consumo de memória com baixo impacto na qualidade das inferências [Dettmers and Zettlemoyer 2023, Lima et al. 2026]. Os experimentos foram executados em uma GPU NVIDIA RTX PRO 6000 Blackwell com 96 GB de VRAM GDDR7.

A avaliação utilizou como *Ground Truth* o conjunto de relatórios técnicos interpretados manualmente pelo pesquisador especialista responsável pela identificação das vulnerabilidades. Todos os LLMs receberam o mesmo *prompt* (comando) de sistema, contendo a definição das categorias Macro e exemplos *Few-Shot*. Como saída, cada modelo produziu uma lista de vulnerabilidades identificadas, acompanhadas de suas evidências, classificação e justificativa.

Como as respostas textuais produzidas pelos LLMs apresentam elevada variabilidade, inviabilizando sua associação direta às amostras do *Ground Truth* por técnicas baseadas em REGEX, adotou-se a abordagem *LLM-as-a-Judge* para realizar essa correspondência. O *Judge* (Juiz) compara os achados produzidos pelos modelos ao *Ground Truth*, classificando cada predição como verdadeira positiva (VP) ou falsa positiva (FP) e associando-a ao respectivo identificador da vulnerabilidade. Essa correspondência também permite identificar falsos negativos (FN), conforme ilustrado na Figura 2.

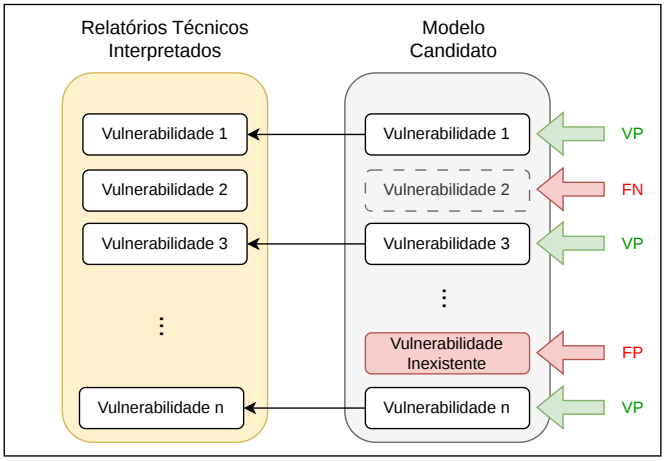

Figura 2. Processo de Avaliação dos Modelos Candidatos

Dessa forma, com a identificação das categorias de erro, é possível calcular as métricas *Precision* (Precisão), *Recall* (Revocação) e *F1-Score* (Medida F1), que são padrões de avaliação em problemas de classificação. A *Precision* mede a proporção de vulnerabilidades corretamente identificadas (VP) entre todas as predições do modelo, enquanto a *Recall* mede a proporção de vulnerabilidades do conjunto *Ground Truth* que foram efetivamente recuperadas.

[4]https://arena.ai/leaderboard/text

O *F1-Score* representa a média harmônica entre *Precision* e *Recall*, será a métrica principal para a avaliação dos modelos candidatos, pois considera de forma equilibrada tanto a capacidade do modelo de encontrar vulnerabilidades reais quanto a de evitar falsos alarmes. As fórmulas de cada métrica são apresentadas a seguir:

$$\text{Precision} = \frac{VP}{VP + FP} \tag{1}$$

$$\text{Recall} = \frac{VP}{VP + FN} \tag{2}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

A metodologia de *LLM-as-a-Judge* representa uma abordagem de estado da arte para avaliação de dados textuais. Contudo, devido à natureza não determinística das LLMs, suas avaliações podem apresentar subjetividade e inconsistências, exigindo procedimentos de calibração e validação do modelo avaliador. Para isso, elaboramos um *dataset* de calibração baseado em uma abordagem *Human-in-the-Loop (HITL)*, composta por rotulações manuais dos achados preditos pelos modelos do *Benchmark*. O especialista pleno em cibersegurança classificou cada vulnerabilidade como verdadeiro ou falso positivo, além de associar os IDs correspondentes às amostras do conjunto *Ground Truth* para cálculo dos falsos negativos. Desse modo, foi possível monitorar o desempenho do LLM *Judge* em suas duas funções principais: (i) rotular vulnerabilidades como corretas ou incorretas; e (ii) atribuir IDs aos apontamentos correspondentes. A primeira afeta o quão representativo será o valor da métrica *Precision*, enquanto que a segunda influencia na métrica *Recall*.

Para atuar como *Judge*, o modelo *Open-Weight* gpt-oss-120b foi selecionado. A escolha desse modelo se deve por seu bom desempenho e pelo fato de sua família arquitetural não estar presente entre os modelos avaliados no *Benchmark*, reduzindo o risco de favorecimento implícito a arquiteturas específicas. Durante todas as avaliações, a temperatura foi definida como zero, com o objetivo de minimizar a variabilidade das avaliações. O *prompt* do *Judge* foi refinado iterativamente com base no *dataset* de calibração. Esse processo teve como objetivo alinhar o comportamento do *Judge* aos critérios das avaliações do especialista pleno e da triagem de vulnerabilidades.

O conjunto de dados utilizado nesse processo possui 318 falhas técnicas preditas, avaliadas manualmente. Durante o processo de seleção dessas vulnerabilidades, utilizamos critérios de preservação da distribuição das cinco categorias Macro, bem como dos modelos utilizados no *Benchmark*. Essas medidas visaram garantir maior representatividade do conjunto de dados de calibração. Os desvios observados entre as métricas produzidas pelo *Judge* e os valores de referência do *dataset* são incorporados como margens de erro nos resultados reportados. Dessa forma, os valores finais das métricas de classificação levam em conta não apenas o desempenho dos modelos, como também a incerteza no processo de avaliação.

Tabela 2. Resultados do Benchmark

Modelo	Tamanho	F1-Score	Recall	Precision	Média CVSS	Sev Crítica
qwen3.6-35b-a3b	35.0	72.0% ± 2.0	60.9% ± 2.2	88.0% ± 1.4	5.28	17
gemma-4-26b-a4b	26.0	68.6% ± 2.0	53.4% ± 2.2	96.2% ± 1.4	5.98	24
qwen3.6-27b	27.0	65.4% ± 2.0	51.8% ± 2.2	88.9% ± 1.4	5.68	17
llama-4-scout-17b-16e-instruct	109.0	61.7% ± 2.0	46.6% ± 2.2	91.3% ± 1.4	5.32	23
gemma-4-e2b	5.1	61.0% ± 2.0	44.7% ± 2.2	96.0% ± 1.4	5.96	23
llama-3.3-70b-instruct	70.0	60.2% ± 2.0	45.1% ± 2.2	90.8% ± 1.4	5.72	24
deepseek-r1-distill-llama-70b	70.0	55.6% ± 2.0	39.1% ± 2.2	95.9% ± 1.4	5.94	22
deepseek-r1-distill-qwen-14b	14.0	54.3% ± 2.0	38.3% ± 2.2	92.9% ± 1.4	5.21	20
deepseek-r1-distill-qwen-32b	32.0	53.5% ± 2.0	37.5% ± 2.2	93.1% ± 1.4	5.76	21
phi-4	14.0	51.8% ± 2.0	35.6% ± 2.2	95.0% ± 1.4	5.75	22
meta-llama-3.1-8b-instruct	8.0	50.3% ± 2.0	36.0% ± 2.2	83.5% ± 1.4	5.95	21
deepseek-r1-distill-llama-8b	8.0	48.2% ± 2.0	34.0% ± 2.2	82.8% ± 1.4	4.95	17
llama-3.2-3b-instruct	3.0	46.6% ± 2.0	35.2% ± 2.2	68.9% ± 1.4	4.80	10
phi-4-mini-instruct	3.8	32.8% ± 2.0	20.6% ± 2.2	81.1% ± 1.4	6.00	13
deepseek-r1-distill-qwen-7b	7.0	27.3% ± 2.0	17.0% ± 2.2	69.7% ± 1.4	5.97	10

5. Experimentos & Resultados

5.1. Benchmark de Modelos

O *Benchmark* avaliou quinze LLMs *Open-Weight* de diferentes famílias e tamanhos. A Tabela 2 apresenta os resultados para *F1-Score*, *Recall*, *Precision*, média do *CVSS Score* e número de vulnerabilidades críticas detectadas. Os valores de *F1-Score*, *Recall* e *Precision* incorporam as margens de erro obtidas durante a calibração do modelo *Judge* (±2,0%, ±2,2% e ±1,4%, respectivamente).

O qwen3.6-35b-a3b apresentou o maior *F1-Score* (72,0%), impulsionado pelo melhor *Recall* do *Benchmark* (60,9%). Em contrapartida, sua *Precision* ocupou apenas a nona colocação (88,0%), indicando um *trade-off* entre capacidade de detecção e ocorrência de falsos positivos. Esse comportamento sugere maior adequação a cenários em que a prioridade é minimizar a omissão de vulnerabilidades.

O gemma-4-26b-a4b obteve o segundo maior *F1-Score* (68,6%), destacando-se pelo melhor resultado em *Precision* (96,2%) e pelo segundo maior *Recall* (53,4%), demonstrando o melhor equilíbrio entre capacidade de detecção e confiabilidade. Além disso, apresentou maior média de CVSS e maior número de vulnerabilidades críticas detectadas que o qwen3.6-35b-a3b, sugerindo melhor desempenho na identificação de falhas de maior severidade no conjunto avaliado.

Embora os modelos anteriores apresentem os melhores resultados do *Benchmark*, seu custo computacional pode limitar a adoção em ambientes com infraestrutura restrita. Nesse contexto, o gemma-4-e2b destacou-se entre os modelos compactos, superando candidatos significativamente maiores, como Llama 70B, DeepSeek-R1 Distill (14B, 32B e 70B) e Phi-4. Apesar do menor *Recall*, apresentou resultados próximos aos do gemma-4-26b-a4b em *Precision*, média de CVSS e identificação de vulnerabilidades críticas, oferecendo um *trade-off* favorável entre desempenho e custo computacional.

5.2. Avaliação com Especialistas

Um experimento adicional desta pesquisa consistiu na comparação direta entre o desempenho humano e os resultados obtidos por um dos melhores modelos LLMs *Open-Weight* classificados no *Benchmark*.

Para isso, a pesquisa contou com o apoio de doze estudantes de pós-graduação em Segurança Cibernética, um grupo de especialistas em formação com conhecimentos qualificados na área, representando uma base tecnicamente capacitada para a execução da tarefa proposta. Durante o experimento, cada participante recebeu individualmente um dos vinte relatórios técnicos utilizados na mesma avaliação aplicada sobre os modelos, realizando manualmente as atividades de interpretação, rotulação e classificação de vulnerabilidades. Para capacitação dos voluntários, foram utilizados como referência os testes de *Usability Assessment Settings*, descritos por [Barros et al. 2026]. A Figura 3 descreve o fluxo experimental dividido em três fases principais.

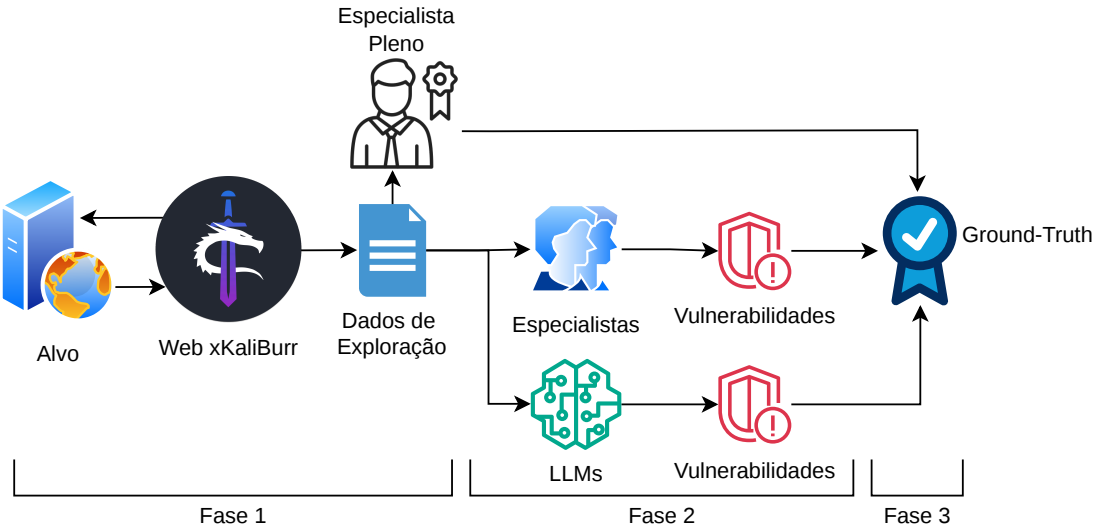

Figura 3. Fluxo do Experimento de Comparação Humano vs. LLM

Na Fase 1, a ferramenta Web xKaliBurr é executada sobre websites e aplicações *online*, realizando as explorações OSINT para coletar informações sobre os ambientes analisados, posteriormente condensadas nos dados de exploração. Nessa mesma etapa, os dados coletados são enviados ao especialista pleno, responsável pela construção do *Ground Truth*, composto pela seleção das vulnerabilidades reais existentes nos ambientes avaliados. Na Fase 2, os dados de exploração são utilizados como insumos para os modelos LLMs *Open-Weight* e para os participantes humanos, que executam atividades de interpretação, detecção e classificação de vulnerabilidades, gerando resultados interpretados contendo descrições das falhas identificadas. Por fim, na Fase 3, os resultados produzidos pelos especialistas humanos e pelos modelos LLMs são comparados ao *Ground Truth*, permitindo medir o desempenho entre as análises.

Após a conclusão da atividade, os participantes responderam a um questionário teórico para relatar suas experiências e dificuldades. Os resultados mostraram que apenas 33,3% possuíam experiência prévia em análise de vulnerabilidades, embora 83,3% já tivessem contato com algumas das falhas descritas nos relatórios técnicos avaliados. Em relação ao esforço empregado, 50% dos voluntários dedicaram entre duas a quatro horas à atividade, enquanto 41,7% relataram mais de quatro horas de trabalho, evidenciando a elevada carga necessária para a realização manual das análises, ao longo dos doze experimentos os participantes identificaram um total de 116 amostras de vulnerabilidades.

Por fim, a Tabela 3 apresenta os resultados das comparações diretas entre as análises manuais realizada pelos especialistas e as inferências produzidas pelo LLM campeão, ao longo das doze rodadas experimentais. O modelo gemma4-26b-a4b foi escolhido para competir com os voluntários humanos por ter apresentado o maior desempenho em *Precision*, descrita como a quantidade de vulnerabilidades corretas encontradas dividida pelo total de vulnerabilidades apontadas, já a comparação de *Recall* representa a quantidade de vulnerabilidades corretas encontradas dividida pelo total de vulnerabilidades existentes no *Ground Truth*.

Tabela 3. Comparação entre Especialistas Humanos e LLM Campeão

Experimento	Quantidade Vulne.		Tempo		Recall		Precision	
	Especialista	Gemma4	Especialista	Gemma4	Especialista	Gemma4	Especialista	Gemma4
1	11/12	7/12	+4h	19.7s	0.92	0.58	0.85	1.0
2	15/15	8/15	2h-4h	23.6s	1.0	0.46	1.0	0.87
3	9/16	6/16	2h-4h	15.5s	0.56	0.31	0.82	0.83
4	10/11	6/11	1h-2h	17s	0.91	0.54	1.0	1.0
5	10/12	6/12	+4h	21.4s	0.83	0.41	0.83	0.83
6	7/15	9/15	2h-4h	22.1s	0.47	0.53	1.0	0.88
7	8/14	5/14	+4h	17.9s	0.57	0.35	1.0	1.0
8	5/12	7/12	2h-4h	20.2s	0.42	0.58	1.0	1.0
9	7/9	5/9	2h-4h	17.7s	0.78	0.44	1.0	0.8
10	6/9	6/9	+4h	19.4s	0.67	0.66	0.86	1.0
11	11/11	8/11	2h-4h	20.8s	1.0	0.63	0.92	0.87
12	4/14	5/14	+4h	21s	0.29	0.35	1.0	1.0

6. Conclusões

De modo geral, os resultados do *Benchmark* demonstram que modelos LLMs *Open-Weight* apresentam potencial significativo para apoiar tarefas de detecção e classificação de vulnerabilidades em atividades de Segurança Cibernética. Os experimentos evidenciaram diferentes perfis de desempenho entre os modelos avaliados, destacando o qwen3.6-35b-a3b pela elevada capacidade de detecção de vulnerabilidades e o gemma-4-26b-a4b pelo equilíbrio entre *Precision*, *Recall* e identificação de falhas críticas. Além disso, os resultados obtidos pelo gemma-4-e2b indicam que modelos menores também podem alcançar desempenho competitivo, oferecendo melhor relação entre custo computacional e eficiência operacional. Em conjunto, os achados reforçam a viabilidade da utilização de LLMs como mecanismos de suporte à análise automatizada de vulnerabilidades, especialmente em cenários que demandam redução do esforço manual e maior escalabilidade nos processos de triagem de vulnerabilidades cibernéticas.

A análise comparativa entre os doze especialistas e o modelo gemma-4-26b-a4b evidencia um cenário de compensação entre qualidade e eficiência. Os analistas humanos apresentaram superioridade em *Recall*, com média de 0.70 contra 0.49 do modelo, vencendo em nove dos doze casos avaliados, enquanto ambos os grupos obtiveram resultados semelhantes em *Precision*, com médias de 0.94 para os especialistas e 0.92 para o modelo.

Entretanto, a maior diferença foi observada no tempo de execução: enquanto os participantes humanos demandaram entre uma a quatro horas ou mais por análise, o modelo concluiu a mesma tarefa em aproximadamente quinze a vinte e quatro segundos, representando uma aceleração estimada entre 300 e 900 vezes. Esses resultados demonstram que, embora os humanos mantenham vantagem na detecção de vulnerabilidades, o modelo oferece uma alternativa viável para cenários em que a velocidade é crítica, atuando também como ferramenta complementar para reduzir a carga de trabalho.

7. Trabalhos Futuros

Como perspectivas futuras, a equipe do projeto propõe o desenvolvimento de um novo modelo LLM *Open-Weight* especializado, construído a partir da combinação das principais características mais vantajosas observadas nos modelos selecionados pelo *Benchmark*, integrando capacidades de detecção, classificação e priorização de vulnerabilidades. A equipe pretende aplicar técnicas avançadas de *Fine-Tuning* e destilação de conhecimento entre os modelos, considerando que pesquisas recentes demonstram que modelos pequenos podem alcançar desempenho semelhante ao de arquiteturas maiores em tarefas específicas por meio desse processo [Caminha et al. 2025]. O objetivo é construir uma solução compacta, eficiente e adequada à integração com ferramentas de exploração OSINT. Paralelamente, o projeto prevê a ampliação do conjunto de *Ground Truth* por meio da colaboração de outros especialistas em cibersegurança, aumentando o número de relatórios técnicos interpretados e fortalecendo a qualidade do processo de rotulação das vulnerabilidades, utilizadas para medir a capacidade dos modelos de IA. Por fim, pesquisas futuras poderão integrar essas estratégias em uma arquitetura mais ampla, capaz de automatizar etapas de detecção, validação e proposição de correções de vulnerabilidades, avançando em direção à automação do fluxo operacional de *Pentest* em ambientes *online*.

Agradecimentos

Este trabalho foi apoiado pela Lenovo, por meio de investimentos em Pesquisa e Desenvolvimento (P&D), realizados em conformidade com a Lei nº 8.248/1991 (Lei de Informática). Os autores agradecem pela contribuição, fundamental para o desenvolvimento desta pesquisa.

Referências

Anderson, R. (2010). *Security engineering: a guide to building dependable distributed systems*. John Wiley & Sons.

Barros, D. R., Cabral, L., Oliveira, J. V., Castro, F. M., Soares, L. L., Monteiro, J. M., Bento, J., and Rocha, L. S. (2024). Web xkaliburr: uma plataforma online para levantamento de informaçoes em pentest em aplicaçoes na internet. In *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais (SBSeg)*, pages 177–184. SBC.

Barros, D. R., Cabral, L., Oliveira, J. V., Castro, F. M., Soares, L. L., Monteiro, J. M., Bento, J., and Rocha, L. S. (2026). Web xkaliburr: An online platform for information gathering in pentest for internet applications. *Journal of the Brazilian Computer Society*, 32(1):700–714.

Bell, D. E. and LaPadula, L. J. (1973). Secure computer systems: Mathematical foundations. Technical Report MTR-2547, MITRE Corporation, Bedford, MA, USA.

Caminha, C., Silva, M. d. L. M., Chaves, I. C., Brito, F. T., Farias, V. A., and Machado, J. C. (2025). DiagHW: A compact LLM for hardware failure diagnosis via a novel knowledge distillation pipeline. In *Brazilian Conference on Intelligent Systems*, pages 393–407. Springer.

Carvalho, R., editor (2023). *OSINT: do zero à investigação profissional*. [s.n.], [S.l.]. eBook Kindle.

Chopra, S., Ahmad, H., Goel, D., and Szabo, C. (2026). Chatnvd: Advancing cybersecurity vulnerability assessment with large language models. *IEEE Access*.

Cuong Nguyen, H., Tariq, S., Baruwal Chhetri, M., and Quoc Vo, B. (2025). Towards effective identification of attack techniques in cyber threat intelligence reports using large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 942–946.

Dettmers, T. and Zettlemoyer, L. (2023). The case for 4-bit precision: k-bit inference scaling laws. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Fielding, R. T. and Reschke, J. (2014). Hypertext transfer protocol (HTTP/1.1): Semantics and content. Technical Report RFC 7231, Internet Engineering Task Force (IETF).

FIRST (2019). Common vulnerability scoring system (CVSS). Acesso em: 01 ago. 2026.

GitHub Security Lab (2025). GitHub security advisories (GHSA). Acesso em: 01 ago. 2026.

Hugging Face (2023). The tokenizer playground. Acesso em: 01 ago. 2026.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., et al. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Lampson, B. W. (1974). Protection. *ACM SIGOPS Operating Systems Review*, 8(1):18–24.

Li, F., Jiang, J., Chen, D., and Xiong, Y. (2026). LLM-based vulnerability detection at project scale: An empirical study. *arXiv preprint arXiv:2601.19239*.

Lima, J. V., Pinheiro, V., and Caminha, C. (2026). Portuguese sentiment analysis with open-source LLMs: Models, prompts, and efficient deployment. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)-Vol. 1*, pages 212–221.

MITRE Corporation (2025). Common Vulnerabilities and Exposures (CVE) and Common Weakness Enumeration (CWE). `https://www.cve.org/`; `https://cwe.mitre.org/`. Acesso em: 01 ago. 2026.

National Institute of Standards and Technology (1995). An introduction to computer security: The nist handbook. Technical Report Special Publication 800-12, NIST.

National Security Agency (2012). Defense in depth: A practical strategy for achieving information assurance in today's highly networked environments. Technical report, National Security Agency (NSA).

OWASP Foundation (2025). OWASP top 10: 2025. Acesso em: 01 ago. 2026.

OWASP LLMs Risk (2025). OWASP Top 10 for Large Language Model Applications. Acesso em: 01 ago. 2026.

Roy, S., Sharmin, N., Acosta, J. C., Kiekintveld, C., and Laszka, A. (2022). Survey and taxonomy of adversarial reconnaissance techniques. *ACM Computing Surveys*, 55(6):1–38.

Theisen, C., Munaiah, N., Al-Zyoud, M., Carver, J. C., Meneely, A., and Williams, L. (2018). Attack surface definitions: A systematic literature review. *Information and Software Technology*, 104:94–103.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*.

World Economic Forum (2025). Global cybersecurity outlook 2025. Technical report, World Economic Forum. Acesso em: 01 ago. 2026.