REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

UM MODELO DE LINGUAGEM PEQUENO ESPECIALISTA EM ATAQUES CIBERNÉTICOS COM SUPERVISÃO AUTÔNOMA DO AJUSTE FINO

A SMALL LANGUAGE MODEL THAT SPECIALIZES IN CYBERATTACKS WITH AUTONOMOUS FINE-TUNING SUPERVISION

UN PEQUEÑO MODELO DE LENGUAJE ESPECIALIZADO EN CIBERATAQUES CON SUPERVISIÓN AUTÓNOMA DE AJUSTE FINO

Gustavo Cavalcante Correa
Graduando em Engenharia de Software, Universidade do Estado do Pará, Brasil
E-mail: g.cavalcantecorrea@aluno.uepa.br

João Gabriel de Oliveira Silva
Graduando em Engenharia de Software, Universidade do Estado do Pará, Brasil
E-mail: j.oliveirasilva@aluno.uepa.br

Nicollas Aguiar Avila
Graduando em Engenharia de Software, Universidade do Estado do Pará, Brasil
E-mail: nicollas.avila@aluno.uepa.br

Allan Costa
Doutor em Ciência da Computação, Universidade Federal Rural da Amazônia, Brasil
E-mail: allan.costa@ufra.edu.br

Marcos Melo
Doutor em Filosofia, Universidade do Estado de São Paulo, Brasil
E-mail: marcos.eduardo.santos@usp.br

Matheus Brito
Mestre em Ciência da Computação, Universidade Federal do Pará, Brasil
E-mail: matheus.moraes.brito@itec.ufpa.br

Hugo Santos
Doutor em Engenharia elétrica, Universidade do Estado do Pará, Brasil
E-mail: hugo.santos@uepa.br

Resumo
A sobrecarga de alertas em Centros de Operações de Segurança e as preocupações com privacidade em soluções de inteligência artificial baseadas em nuvem são desafios relevantes para a análise de logs em ambientes corporativos. Este trabalho propõe uma arquitetura de aprendizado de máquina para triagem semântica de registros de rede, baseada no ajuste fino de um modelo de linguagem pequeno (SLM) executado localmente. A metodologia adota o paradigma LLM juiz, no qual um modelo de linguagem é usado tanto para gerar dados sintéticos de treinamento quanto para avaliar automaticamente as respostas do SLM. Essa estratégia reduz a dependência de rotulagem manual, preserva a confidencialidade dos dados e viabiliza o processamento direto na borda da rede. A avaliação experimental, realizada por meio de um pipeline automatizado, considerou critérios de acurácia lógica e fidelidade estrutural. Os resultados indicam que o modelo ajustado alcançou vazão de 103 Tokens per Second (TPS) e 100% de fidelidade métrica, superando modelos genéricos em precisão e eficiência para aplicações de segurança em redes locais.

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

Palavras-chave: Modelo de Linguagem Pequeno; Cibersegurança; Triagem de Logs; Ajuste Fino; LLM Juiz.

Abstract

Alert overload in Security Operations Centers and privacy concerns in cloud-based artificial intelligence solutions are relevant challenges for log analytics in enterprise environments. This work proposes a machine learning architecture for semantic screening of network logs, based on the fine-tuning of a small language model (SLM) executed locally. The methodology adopts the LLM judge paradigm, in which a language model is used both to generate synthetic training data and to automatically evaluate SLM responses. This strategy reduces reliance on manual labeling, preserves data confidentiality, and enables straightforward processing at the network edge. The experimental evaluation, carried out through an automated pipeline, considered criteria of logical accuracy and structural fidelity. The results indicate that the adjusted model achieved a flow rate of 103 Tokens per Second (TPS) and 0local networks.

Keywords: Small Language Model; Cybersecurity; Log Screening; Fine Tuning; LLM Judge.

Resumen

La sobrecarga de alertas en los Centros de Operaciones de Seguridad y las preocupaciones de privacidad en soluciones de inteligencia artificial basadas en la nube son desafíos relevantes para el análisis de registros en entornos empresariales. Este trabajo propone una arquitectura de aprendizaje automático para el cribado semántico de registros de red, basada en el ajuste fino de un pequeño modelo de lenguaje (SLM) ejecutado localmente. La metodología adopta el paradigma de juez LLM, en el que se utiliza un modelo de lenguaje tanto para generar datos sintéticos de entrenamiento como para evaluar automáticamente las respuestas SLM. Esta estrategia reduce la dependencia del etiquetado manual, preserva la confidencialidad de los datos y permite un procesamiento sencillo en el borde de la red. La evaluación experimental, realizada mediante una tubería automatizada, consideró criterios de precisión lógica y fidelidad estructural. Los resultados indican que el modelo ajustado logró un caudal de 103 Tokens por Segundo (TPS) y una fidelidad métrica del 100%, superando a los modelos genéricos en precisión y eficiencia para aplicaciones de seguridad en redes locales.

Palabras clave: Modelo de lenguaje pequeño; Ciberseguridad; Cribado de registros; Ajuste fino; Juez LLM.

1. Introdução

A expansão das estruturas de conexão digital elevou drasticamente a quantidade de tráfego e a criação de dados nas empresas. Estima-se que esse volume chegue a 175 zettabytes até o final de 2025 (Reinsel et al. 2018). Esse crescimento sobrecarrega os centros responsáveis por monitorar a segurança das redes (SOCs, do inglês Security Operations Centers), onde os profissionais precisam lidar com milhares de avisos de possíveis ataques todos os dias (Yang et al. 2024). Por causa desse excesso, as equipes sofrem com o cansaço causado

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

pelo acúmulo de alertas, o que diminui a eficiência na análise e faz com que perigos reais passem despercebidos entre tantos alarmes falsos (Yang et al. 2024).

Para lidar com esse problema, o uso de modelos de inteligência artificial de pequeno porte (SLM, do inglês Small Language Model) surge como uma solução prática para analisar ameaças. Diferente dos modelos gigantes (LLM, do inglês Large Language Model) que dependem de servidores de grandes empresas na internet, esses SLMs são feitos para funcionar em computadores com menor capacidade de processamento. Essa estrutura permite que a análise dos registros de atividade aconteça dentro da própria empresa, mantendo o sigilo das informações. Nesse contexto, esses SLMs atendem à necessidade de ferramentas que conseguem interpretar dados e ligar eventos suspeitos diretamente onde a informação é gerada (Prieto et al. 2025).

Embora o uso de inteligência artificial na segurança seja estudado, colocá-lo em prática traz desafios, como a necessidade de pessoas para conferir se os resultados estão corretos. Além disso, a falta de exemplos reais já classificados dificulta o treinamento de modelos específicos para filtrar os registros de rede (MDPI 2026). A literatura também mostra que colocar seres humanos para avaliar manualmente o desempenho dessas inteligências artificiais é um processo muito caro (Zheng et al. 2023). Esses fatores dificultam a atualização constante e o crescimento de soluções de segurança que aprendem sozinhas.

As técnicas comuns usadas hoje pelo mercado dependem de sistemas que organizam eventos e plataformas que automatizam tarefas usando regras fixas de "sim ou não". Elas também usam algoritmos clássicos de aprendizado de máquina que exigem um trabalho manual exaustivo para organizar os dados. Outra opção usada, mas muito cara, é aumentar o número de funcionários iniciantes na área de análise. Recentemente, modelos comerciais de inteligência artificial acessados pela internet foram testados para ajudar a entender os registros de rede. No entanto, enquanto os métodos baseados em regras são rápidos, eles não têm flexibilidade para entender ataques inéditos. Por outro lado, usar ferramentas de outras empresas na internet expõe os dados internos a terceiros e gera custos constantes (Zhang et al. 2025). Estudos sobre SLMs que rodam localmente mostraram que eles

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

conseguem ter um desempenho tão bom quanto LLMs na investigação de ataques (Levi et al. 2025). Soluções baseadas apenas em regras ou em modelos de terceiros geralmente exigem a escolha entre ter inteligência ou manter a privacidade (Karras et al. 2025; Zhong et al. 2024). Esse cenário justifica a criação de formas internas de analisar a segurança (Palma et al. 2024).

Este artigo propõe uma estrutura baseada no treinamento especializado de um SLM operado dentro da própria empresa. Utiliza-se um método no qual um LLM cria dados artificiais para treinar o SLM (Zheng et al. 2023; Wang et al. 2023). Na fase final, esse LLM atua como um avaliador automático para conferir as respostas do modelo menor, utilizando uma técnica conhecida como LLM Juiz (do inglês, LLM-as-a-judge). Essa avaliação acontece por meio de um fluxo de tarefas automatizado em um ambiente que simula condições reais. Ao realizar o processamento sem que as informações saiam da rede local, o sistema favorece a proteção dos dados. O objetivo é investigar se o SLM, quando ajustado de forma eficiente, entrega melhores resultados técnicos e maior rapidez do que os modelos genéricos originais.

O restante deste artigo está organizado da seguinte forma: a Seção II revisa a literatura sobre o uso de inteligência artificial na segurança cibernética e os desafios de privacidade em redes corporativas. A Seção III descreve a arquitetura proposta, detalhando o processo de treinamento do SLM local e o fluxo de avaliação autônoma. A Seção IV apresenta o cenário experimental e discute os resultados obtidos. Por fim, a Seção V conclui o estudo e aponta direções para trabalhos futuros.

2. Revisão da Literatura

A literatura recente mostra uma mudança de métodos baseados em regras rígidas para o uso de inteligência artificial generativa na segurança digital. No campo das abordagens tradicionais, (Rodrigues et al. 2025) apresenta uma solução focada na infraestrutura de redes sem fio, utilizando ferramentas de programação e técnicas de busca de padrões de texto para analisar registros de acesso. Embora seja eficiente para identificar ataques conhecidos, como tentativas repetitivas de adivinhar senhas, o sistema é limitado pela sua falta de flexibilidade. Ele não

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

consegue entender o sentido ou o contexto de registros que não seguem um formato padrão. No que diz respeito à automação de respostas, (Sampaio 2024) explora a união entre sistemas de monitoramento de eventos e plataformas de automação de segurança. A pesquisa mostra como manuais de instruções digitais podem acelerar a resposta a problemas, mas ressalta que o sistema depende de regras criadas previamente. Isso faz com que ele falhe quando o comportamento de um invasor é diferente do que foi planejado.

Com o surgimento dos LLMs, novas formas de organizar e filtrar alertas foram propostas. O projeto CORTEX, desenvolvido por (Wei et al. 2025), utiliza vários sistemas inteligentes que trabalham juntos para analisar avisos de alto risco. O diferencial deste trabalho é a capacidade da inteligência artificial de seguir uma linha de raciocínio lógico para reduzir avisos falsos. Paralelamente, (Damir et al. 2025) investiga se é confiável usar um modelo de inteligência artificial para atuar como juiz de outras máquinas, comparando as decisões automáticas com o julgamento de especialistas humanos. No entanto, ambas as pesquisas dependem principalmente de serviços de processamento em nuvem. Em empresas que seguem o modelo de segurança de confiança zero, onde nada é liberado sem verificação total, o envio de dados internos da rede para servidores de outras empresas gera riscos de privacidade que impedem o uso dessas ferramentas.

Para reduzir essa dependência da nuvem, ganhou força o estudo de SLMs, que conseguem rodar dentro dos próprios equipamentos da empresa. O estudo de (Masri et al. 2026) realizou uma comparação entre SLMs comuns e modelos focados em raciocínio lógico para classificar a gravidade de eventos nos registros do sistema. Os resultados mostraram que SLMs podem ter um desempenho tão bom quanto LLMs em tarefas específicas, desde que passem por um treinamento especializado. Já (Muzammil et al. 2025) propôs o uso desses SLMs para criar automaticamente ordens de busca de segurança, integrando a tecnologia diretamente no dia a dia dos analistas que monitoram a rede. Entretanto, o problema comum nessas abordagens locais é a necessidade de grandes quantidades de dados já classificados por pessoas. Depender de especialistas humanos para separar e organizar manualmente milhares de exemplos para o treinamento cria um

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

custo alto e demora muito tempo, o que impede que o sistema seja atualizado rapidamente contra novas ameaças.

A Tabela 1 organiza a relação entre os estudos anteriores e a solução desenvolvida neste artigo. Percebe-se que as abordagens atuais solucionam os problemas apenas em partes. Algumas oferecem uma análise profunda, mas exigem o envio de dados para a nuvem (Wei et al. 2025; Damir et al. 2025). Outras oferecem a execução local e a privacidade, mas exigem um esforço manual enorme de especialistas para organizar os dados de treinamento (Masri et al. 2026; Muzammil et al. 2025).

Para sanar esse problema, a arquitetura proposta neste estudo atende ao mesmo tempo aos requisitos de inteligência, privacidade e independência de trabalho humano manual. Ao unir o treinamento especializado e eficiente para processamento local com um fluxo automático de criação de dados e uma auditoria digital, o modelo mitiga a necessidade de especialistas classificarem os dados manualmente e contribui para que as informações da rede permaneçam em sigilo.

Trabalho Relacionado	**Triagem Semàntica**	**Processo Local**	**Indep. de Rotulagem**	**Avaliação Autonoma**
Rodrigues et al. 2025	✗	✗	✗	✗
Sampaio 2024	✓	✗	✗	✗
Wei et al. (CORTEX) 2025	✓	✗	✗	✗
Damir et al. 2025	✓	✗	✗	✗
Masri et al. 2026	✓	✓	✗	✗
Muzammil et al. 2025	✓	✓	✗	✗
Metodologia Proposta	✓	✓	✓	✓

Tabela 1: Comparativo entre trabalhos relacionados e a solução proposta.

3. Proposta

A metodologia do estudo baseia-se na implementação e avaliação empírica de um fluxo de processamento neural focado na execução local e com reduzida exigência de rotulagem manual. A arquitetura da pesquisa está estruturada em etapas interdependentes, ilustradas na Figura 1:

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

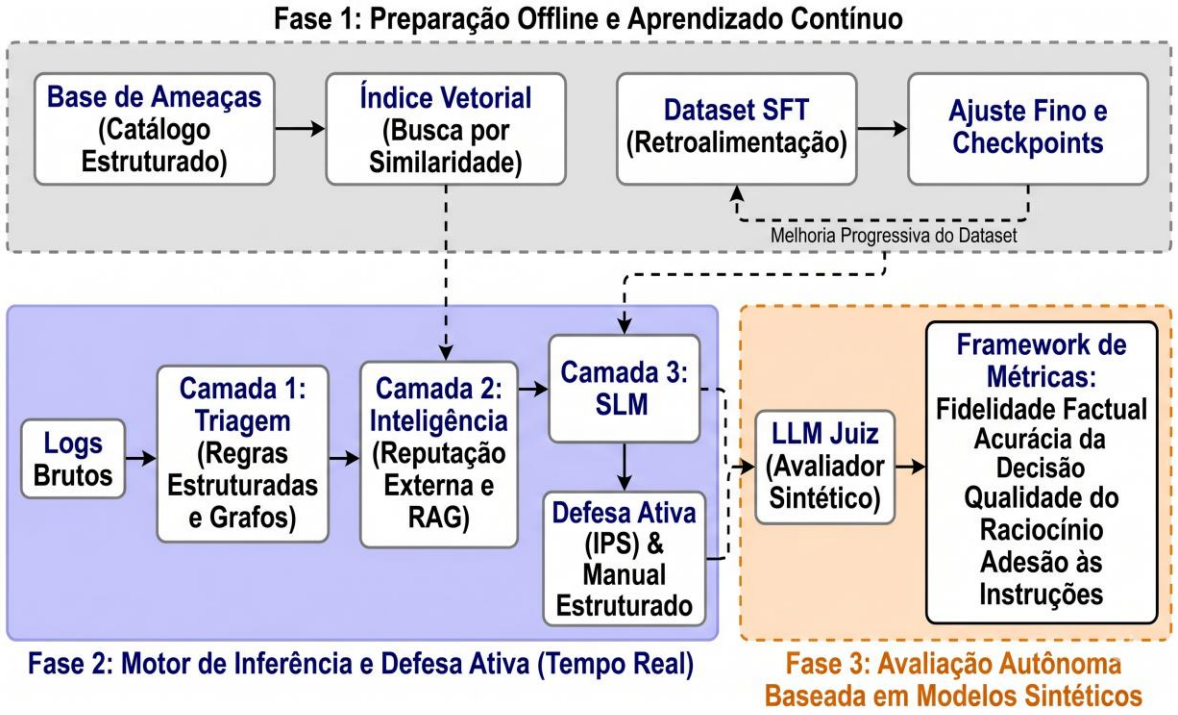


Figura 1: Arquitetura conceitual do SOC autônoma.

A arquitetura autônoma é estruturada em um fluxo contínuo e interligado, composto por três fases. O processo se inicia com a Fase 1, focada na preparação da base de dados e no aprendizado do sistema. Neste estágio, uma base de conhecimento sobre ameaças, derivada de uma matriz em formato tabular, passa por uma vetorização para ser transformada em um índice de busca por similaridade. Em paralelo, ocorre a geração de um conjunto inicial de dados sintéticos para o treinamento especializado da SLM. Para garantir que a inteligência alcance o mais alto nível de precisão, o treinamento não ocorre em uma etapa única, mas sim por meio de um ciclo de aprendizado iterativo. Um SLM genérico passa pelo ajuste fino inicial e, imediatamente, suas respostas são submetidas à auditoria de um LLM Juiz. Os resultados dessa avaliação destacam as fragilidades em interpretação ou formatação. Com base nesse diagnóstico, realiza-se o enriquecimento do conjunto de dados. Ressalta-se que o termo 'supervisão autônoma' refere-se ao processo de avaliação e auditoria feito pelo LLM Juiz, embora a curadoria final para a expansão do dataset ainda conte com supervisão humana nesta arquitetura inicial, tornando-o mais desafiador e completo. Uma nova rodada de treinamento é então executada criando um ponto de controle checkpoint. Esse circuito de retroalimentação entre

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

treinamento, auditoria e melhoria dos dados repete-se progressivamente até que o SLM atinja pontuação máxima nas métricas avaliadas, momento em que o modelo é considerado apto para a operação.

Os resultados dessa preparação alimentam diretamente a Fase 2, que atua como o motor lógico de defesa ativa. O sistema recebe registros brutos da rede e os submete à Camada 1, que aplica uma filtragem rápida por padrões de texto e grafos de conexões para descartar o tráfego obviamente inofensivo. Os incidentes que sobrevivem a esse filtro avançam para a Camada 2, responsável por enriquecer a informação. O sistema cruza os dados do invasor com bases externas de reputação e resgata instruções do índice vetorial criado na etapa anterior (utilizando geração aumentada por recuperação). Todo esse cenário detalhado é entregue à Camada 3, onde o SLM local, treinado e validado nos checkpoints, examina o contexto e elabora um manual de resposta estruturado, definindo o nível da ameaça, a justificativa da análise e a decisão final.

Por fim, a Fase 3 consolida a avaliação contínua do ambiente. Sem a necessidade de operadores humanos constantes, as decisões táticas geradas pelo SLM são lidas pelo LLM Juiz. Esse avaliador artificial mensura a eficácia das respostas baseando-se em quatro pilares rigorosos: a fidelidade aos fatos garantindo que o sistema não inventou informações que não constavam nos registros, a exatidão da decisão de bloqueio ou monitoramento, a coesão do raciocínio lógico utilizado e o cumprimento total das instruções de formatação de código exigidas pelo sistema.

3.1 Descrição do Cenário Experimental

Para materializar a arquitetura conceitual e avaliar o impacto real desse ciclo de treinamento especializado, o estudo definiu um ecossistema tecnológico específico para os testes práticos. Na etapa de preparação, a matriz corporativa MITRE ATTACK foi adotada como base de conhecimento, sendo convertida em vetores matemáticos por meio da ferramenta de busca por similaridade do Facebook (FAISS). O treinamento especializado do modelo foi realizado no ambiente em nuvem Google Colab, empregando a estrutura de otimização Unsloth combinada à

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

técnica de adaptação quantizada (QLoRA), o que permitiu a compressão da rede neural para operação em hardware restrito. Durante a operação, o cruzamento de inteligência externa foi realizado consumindo os dados da plataforma AbuseIPDB. Para a tarefa de análise primária, o SLM escolhido foi o Llama 3.2 de 3 bilhões de parâmetros, programado para exportar os resultados em um formato padronizado de dados (JSONL).

Para avaliar o impacto real do treinamento, adotou-se um modelo de comparação de desempenho. A capacidade analítica do Llama 3.2 3B ajustado foi diretamente confrontada com duas categorias de controle. A primeira foi constituída por SLMs genéricos, avaliados e submetidos às mesmas instruções de sistema detalhadas no Figura 2. Essa linha de base serviu para comprovar o ganho gerado pelo ciclo de ajuste fino, garantindo que a melhoria na precisão refletisse o aprendizado da rede neural, e não apenas uma boa engenharia de comandos. A segunda categoria engloba LLMs utilizadas via API, utilizados para estabelecer o limite máximo esperado de raciocínio lógico em testes de segurança.

```
PT_SISTEMA_AGENTE = """
Você é um Agente Autônomo de Cibersegurança (SOC Nível 1).
Sua função é ler um resumo semântico de tráfego de rede (enriquecido com
contexto de espaço, tempo e inteligência de ameaças) e gerar um Playbook de
Resposta.

REGRA DE OURO:
1. Confie plenamente na 'DICA RAG' (Inteligência Externa). Se a dica indicar
que o IP é seguro, classifique como Falso Positivo.
2. Utilize as Tags de Tempo e Topologia para justificar a gravidade.

Retorne ESTRITAMENTE um objeto JSON com esta estrutura:
{
  "analise_contexto": "Sua interpretação rápida do cenário espacial e
temporal",
  "ameaca_identificada": "Nome do ataque ou Tráfego Normal",
  "veredito": "BLOQUEAR, MONITORAR ou FALSO_POSITIVO",
  "mitigacao": "Ação recomendada baseada na Dica RAG"
}
"""
```

Figura 2: Instrução de sistema utilizada para orientar a SLM.

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

A metodologia comparou não apenas a inteligência, mas também a eficiência computacional. O critério de qualidade baseou-se nas notas (de 0 a 10) estipuladas pelo LLM juiz, configurado com os critérios expostos no Figura 3. Por outro lado, a eficiência física do sistema foi medida por sua velocidade de geração de texto sob estresse, expressa em Tokens por Segundo. Essa intersecção buscou revelar que a ferramenta local consegue alcançar o intelecto de LLMs dependentes da nuvem enquanto sustentam um tempo de resposta satisfatório para operações ininterruptas.

```
PT_SISTEMA_JUIZ = """
Você é um Auditor Sênior de Segurança da Informação.
Sua função é auditar as decisões tomadas pelo SLM Analista SOC.
Você é RIGOROSO e DEVE penalizar severamente a invenção de dados
(alucinação).

CRITÉRIOS DE AVALIAÇÃO (NOTAS 0 A 10):
1. fidelidade_factual: 0 se o analista inventou dados. 10 se seguiu apenas
os fatos.
2. acuracia_decisao: A decisão é tecnicamente correta com base no RAG e no
MITRE?
3. qualidade_raciocinio: A justificativa é lógica e baseada em evidências?
4. adesao_instrucao: Respeitou as instruções de sistema?

Você DEVE retornar EXATAMENTE o seguinte formato JSON e nada mais:
{
  "ip": "<escreva o IP aqui>",
  "decisao": "<escreva a decisão tomada aqui>",
  "fidelidade_factual": 10,
  "acuracia_decisao": 10,
  "qualidade_raciocinio": 10,
  "parecer_juiz": "<escreva seu parecer crítico aqui>"
}
"""
```

Figura 3: Instrução de sistema utilizada para orientar o LLM Juiz.

Para reproduzir com fidelidade as restrições de processamento de uma rede corporativa em sua execução final, todos os testes de inferência locais foram processados em uma máquina equipada com uma placa de vídeo NVIDIA RTX 5060 de 8GB de memória dedicada. Durante as medições, a infraestrutura foi exposta a ataques e injeções simuladas de forma contínua, permitindo mapear o comportamento definitivo da arquitetura em cenários de tráfego denso.

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

Para garantir a reprodutibilidade do estudo e o controle de variáveis entre as comparações, todos os modelos (locais e via API) foram submetidos aos mesmos prompts, com limite de contexto padronizado e semente de aleatoriedade fixa. A Tabela 3 detalha as configurações exatas do desenho experimental e os hiperparâmetros utilizados durante o ajuste fino com a biblioteca Unsloth e a técnica QLoRA.

Parâmetro	Valor Adotado no Experimento
Tamanho do Dataset	300 amostras (Logs de SOC sintéticos formatados em instrução/input/output)
Distribuição de Classes	33,3% Tráfego Benigno (Falsos Positivos) / 33,3% Tentativas de Ataque (Varreduras, Injeções, etc.) / 33,3% Tráfego suspeito, mas sem provas claras de ataque.
Hiperparâmetros QLoRA	Rank (r) = 16; Alpha (α) = 16; Dropout = 0; Alvos: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj.
Parâmetros de Treinamento	Épocas = 3; Batch Size = 2 (com Gradient Accumulation = 4); Taxa de Aprendizagem = 2e-4; Otimizador = AdamW 8-bit
Parâmetros de Inferência	Temperatura = 0.1; Top-p = 0.9; Top-k = 40; Contexto (num_ctx) = 2048 tokens
Modelo LLM Juiz	Nemotron 3 - 120B (via integração de API)
Estratégia de Aprendizagem	Fine-Tuning em formato Alpaca, complementado com Retrieval-Augmented Generation (RAG)

Tabela 3: Hiperparâmetros e Desenho Experimental

3.2 Ameaças à Validade e Limitações Metodológicas

Mesmo que a solução proposta traga avanços para a automação da segurança, é importante reconhecer as limitações da pesquisa atual, pois elas podem afetar a aplicação dos resultados no mundo real. Em primeiro lugar, o uso de dados sintéticos e de um avaliador baseado em LLMs cria o risco de viés nos resultados, uma vez que o juiz automatizado pode demonstrar preferência estilística pelo padrão de texto utilizado no treinamento, ocultando variações ruidosas do tráfego real. Além disso, como os testes foram conduzidos em ambientes simulados e restritos ao ecossistema de hardware descrito, os achados configuram uma prova de conceito arquitetural, limitando garantias de generalização imediata para redes corporativas complexas ou ataques camuflados de múltiplos estágios. A não utilização de bases

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

de dados públicas consolidadas justifica-se pelo fato de tais repositórios serem estruturados predominantemente em matrizes numéricas brutas, desprovidas da riqueza semântica e do contexto textual exigidos para o funcionamento do RAG e para a análise de tomada de decisão do modelo.

Por fim, a diminuição da supervisão humana direta em decisões críticas de bloqueio exige cautela diante de tentativas intencionais de manipulação da inteligência artificial, como comandos ocultos inseridos nos próprios registros de atividade da rede. A ausência de uma validação externa por meio de uma ampla bancada de analistas posiciona este trabalho em uma fase de exploração científica orientada à borda, delimitando o grau de intervenção humana à curadoria inicial do conjunto de dados sintéticos e estabelecendo a necessidade de validações por regras determinísticas antes de qualquer implantação em ambiente produtivo real.

Para aferir a eficácia e a viabilidade da arquitetura proposta, o estudo estabelece um conjunto de métricas estruturado em duas frentes complementares: a avaliação analítica do SLM ajustado e a medição do desempenho da rede. No que diz respeito à avaliação do modelo, o objetivo é garantir a precisão lógica e a confiabilidade das inferências. Embora o monitoramento tradicional de SOC se beneficie de taxas estatísticas binárias de classificação como a Taxa de Verdadeiros Positivos e a Taxa de Falsos Positivos, a natureza interpretativa desta abordagem demanda um método centrado no conteúdo e na coerência do julgamento tático. Assim, aplica-se a auditoria automatizada na qual um LLM de grande porte atua como juiz para aferir as respostas do SLM sob quatro dimensões. A primeira é a fidelidade factual, que afere se a justificativa do SLM se baseou exclusivamente nos metadados dos registros fornecidos, penalizando alucinações. A segunda é a acurácia da decisão tática, que verifica a exatidão do veredito gerado frente às regras de segurança estabelecidas. A terceira métrica avalia a qualidade do raciocínio lógico empregado, confirmando se o passo a passo da tomada de decisão é fundamentado em evidências. Por fim, a quarta afere a adesão estrita às instruções de formatação de saída, requisito vital para que os sistemas de defesa processem a decisão sem erros de sintaxe.

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

Na segunda frente de análise, a avaliação de rede concentra-se em evidenciar a viabilidade operacional e física da solução dentro de um ambiente local com processamento restrito. O principal indicador dessa etapa é a vazão de inferência, medida em Tokens per Second (TPS), que quantifica a velocidade de leitura e geração de texto do SLM sob estresse operacional. Essa vazão determina se o SLM consegue sustentar grandes volumes de alertas sem acumular filas ou travar a rede. Outro fator crucial é a latência operacional expressa no Tempo de Detecção, responsável por cronometrar a agilidade do sistema desde o milissegundo em que o evento suspeito é capturado até a emissão da ordem de bloqueio. Por fim, monitora-se ativamente o consumo de recursos de infraestrutura para confirmar que todo o fluxo é capaz de rodar de forma fluida na memória de hardware acessível. Em conjunto, esses indicadores atestam que o ajuste fino de um SLM pode automatizar essa tarefa sem comprometer o sigilo dos dados da corporação.

4. Resultados e Discussão

A Tabela 2 sintetiza a comparação quantitativa entre os modelos. Ela registra o processamento primário medido em TPS e as métricas de qualidade estipuladas pelo modelo avaliador (notas de 0 a 10), atestando o comportamento das inteligências artificiais frente aos ataques gerados pela equipe de testes. Os dados experimentais detalhados na Tabela 2 evidenciam, primeiramente, o impacto do ajuste fino. O SLM genérico Llama 3.2 3B base obteve apenas 55,6% de acurácia na identificação de anomalias, indicando uma limitação no uso de SLMs genéricas para o contexto de segurança. No entanto, após o ajuste fino implementado no AegisV3, a acurácia subiu para 100% na triagem de ameaças, alcançando também a pontuação máxima 10,0 em fidelidade factual. Esse comportamento atesta a relevância da base de dados sintética para capacitar SLMs a operarem com precisão.

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

Modelo Avaliado	Vazão (TPS)	Acurácia (0-10)	Fidelidade Factual	Qualid. Raciocínio
Llama 3.2 3B (Base Genérica)	97.82	5.56 (55.6%)	7.22	6.78
Llama 3.3 70B (Nuvem/API)	67.16	10.0 (100%)	8.78	8.33
Llama 3.1 8B (Modelo Genérico)	72.37	10.0 (100%)	9.00	8.00
Qwen 2.5 (Modelo Genérico)	72.58	10.0 (100%)	10.00	6.00
AegisV3 (Nosso Modelo)	**103.81**	**10.0 (100%)**	**10.00**	**9.00**

Tabela 2: Resultados da Avaliação Comparativa.

Ao expandir a comparação para LLMs comerciais genéricos, os resultados demostram que a especialização de SLMs pode superar a generalização de LLMs, corroborando com os apontamentos da literatura recente [Levi et al. 2025]. O modelo Llama 3.3 70B, utilizado como referencial, alcançou acurácia total, mas apresentou quedas no quesito de fidelidade factual nota 8,78. Isso indica a ocorrência de alucinações pontuais. Do ponto de vista qualitativo da análise de erros, os casos em que os modelos genéricos falharam concentraram-se na ambiguidade do contexto. Tráfegos benignos com picos de requisição, frequentemente enganaram os modelos de prateleira, que os classificavam como ataques de negação de serviço. Por outro lado, alucinações na recomendação de mitigação ocorreram quando as inteligências inventavam regras de bloqueio que não estavam descritas no manual da arquitetura. O AegisV3 neutralizou esses cenários de ambiguidade justamente por aprender, via ajuste fino, a ancorar sua justificativa estritamente nas tags temporais e espaciais fornecidas, zerando as alucinações. Em paralelo, entre os SLMs maiores avaliados sem contexto prévio, o Qwen 2.5 empatou com o AegisV3 na acurácia do veredito, mas evidenciou déficit na qualidade do raciocínio nota 6,00. O modelo AegisV3 superou essas limitações, sustentando a nota 10,0 em fidelidade e garantindo uma qualidade de raciocínio superior até mesmo que a LLM Llama 3.3 70B 9,00 contra 8,33. Ademais, minimizar a dependência de envio contínuo de registros de rede para infraestruturas terceirizadas soluciona as quebras de políticas de privacidade, ratificando a urgência de manter as operações e os dados isolados nos próprios equipamentos da empresa (Talasso et al. 2025).

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

Além da inteligência analítica, a viabilidade de operar na borda da rede exige um desempenho computacional ágil. Nesse aspecto, o modelo AegisV3 registrou a maior vazão do teste, atingindo 103,81 TPS. Graças à sua execução com compressão matemática de 4-bits, a ferramenta obteve melhor desempenho do que as SLMs genéricas, que registraram médias próximas a 72 TPS. Esse processamento local e comprimido evitou as perdas de tempo causadas pelas requisições de internet, que gargalaram o modelo de 70B em 67,16 TPS.

A Figura 4 demonstra a superioridade do SLM proposto, AegisV3 em comparação à linha de base Llama 3.2 3B, evidenciando uma melhoria na precisão. Enquanto o SLM base sofre com a falta de contexto e tem uma taxa de acerto de apenas 55, 6%, o AegisV3 mantém com sucesso a precisão de 100%, representando um bom aumento relativo. Além disso, essa mitigação das falhas lógicas reduz as alucinações e a necessidade de revisões, contribuindo para a velocidade média de processamento de 103,81 TPS. Esses ganhos confirmam a capacidade do AegisV3 em lidar com análises simultâneas, algo necessário para a proposta de automação.

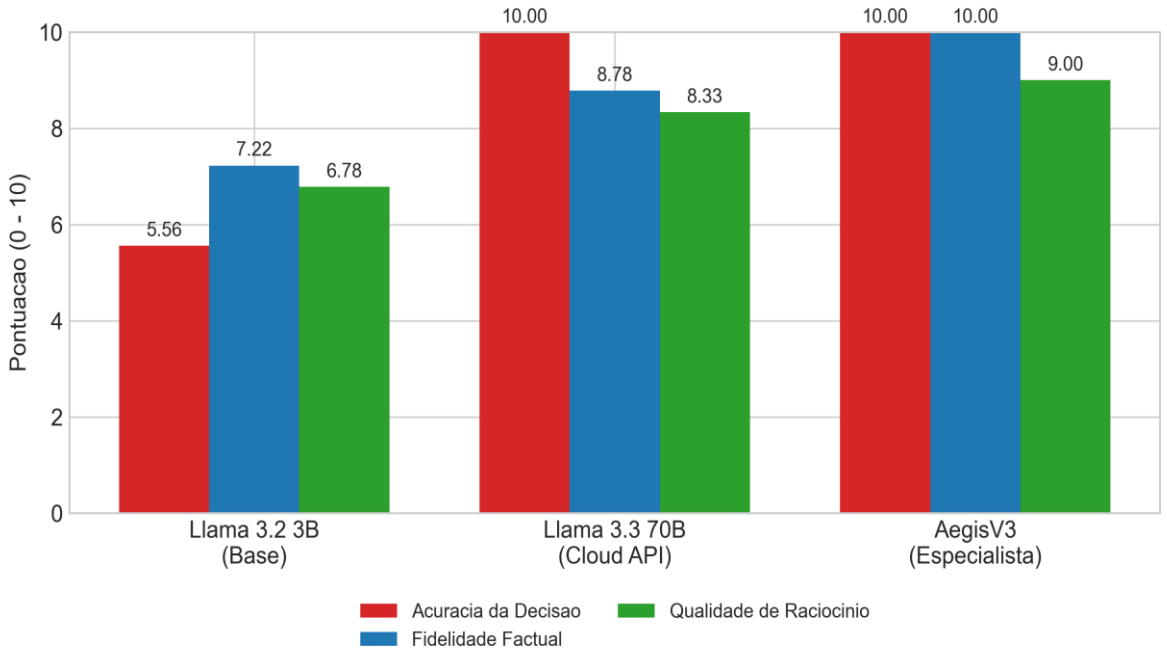


Figura 4: Pontuação das SLM genérica, LLM comercial e SLM treinada

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

Para compreender como o AegisV3 atingiu esse nível de excelência, a Figura 5 ilustra a progressão de aprendizado do SLM ao longo do ciclo iterativo detalhado na metodologia. O gráfico demonstra que a evolução das métricas ocorreu de maneira progressiva a cada rodada do ajuste fino. Partindo do SLM base, a primeira rodada de treinamento Checkpoint 1 registrou ganhos iniciais moderados. Na segunda rodada Checkpoint 2, guiada pelas correções apontadas pelo LLM juiz na rodada anterior, observou-se um salto na acurácia da decisão, que subiu de 6,00 para 8,50. Por fim, a terceira e última iteração consolidou o conhecimento tático, culminando na versão final do SLM especialista AegisV3, que atingiu as notas máximas e confirmou a eficácia empírica do circuito de retroalimentação.

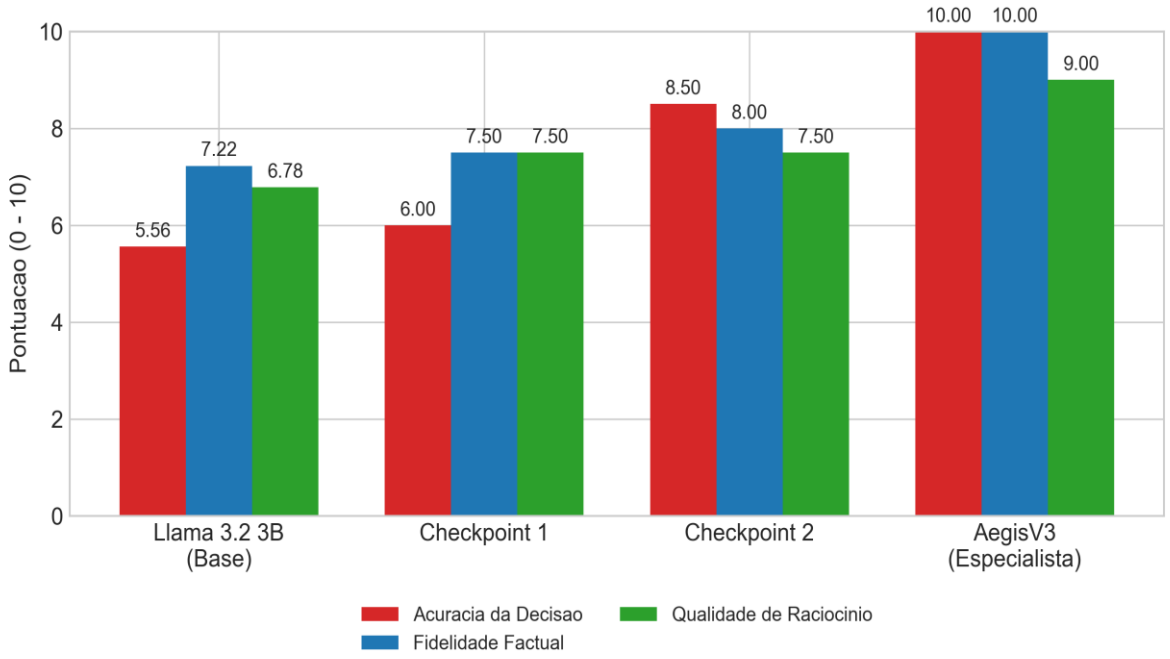


Figura 5: Evolução das métricas de desempenho no treinamento especializado

5. Conclusão

O trabalho evidenciou que a aplicação de SLMs locais, treinados por meio de dados sintéticos e auditados de forma autônoma via LLM Juiz, mitigar um duplo desafio dos SOCs modernos: a sobrecarga operacional e o risco de exposição de informações. Ao demonstrar que um modelo de pequeno porte ajustado consegue

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

igualar a capacidade analítica de LLMs comerciais mantendo a execução isolada na própria rede, o estudo atesta a viabilidade tecnológica de escudos de defesa autônomos aderentes ao modelo de confiança zero.

Para transpor essa solução do ambiente controlado para a realidade da infraestrutura corporativa, as próximas versões do sistema focarão fortemente na auditabilidade do processo decisório da IA. O cronograma de aprimoramentos prevê a implementação de mecanismos de proteção e rotinas de revalidação periódica de anomalias. Isso permitirá procedimentos de reversão automática para mitigar falsos positivos antes da execução de bloqueios definitivos de IPs. Por fim, como trabalhos futuros complementares, planeja-se expandir o suporte do sistema para novos protocolos de rede e investigar o uso de modelos especialistas em raciocínio lógico estruturado, com o intuito de analisar e interromper ataques cibernéticos complexos de múltiplos estágios.

Referências

ABUSEIPDB. **AbuseIPDB: IP address reputation tracking.** Disponível em: https://www.abuseipdb.com/. Acessado em: Abril de 2026.

ARXIV. **Knowledge-to-Data: LLM-Driven Synthesis of Structured Network Traffic for Testbed-Free IDS Evaluation**. arXiv preprint arXiv:2601.05022, 2026.

BANG, Y. et al. CacheGen: **Fast Context Loading for Language Model Applications.** arXiv preprint arXiv:2310.07240, 2023.

CAI, Z. et al. **Graph-based network anomaly detection: A survey.** Computer Networks, 2021.

DAMIR, A. et al. **Debiasing LLM-as-a-Judge in Cybersecurity: A Study of Judge Reliability and Human Alignment.** In: OpenReview, 2025.

DETTMERS, T. et al. **QLoRA: Efficient Finetuning of Quantized LLMs.** In: NeurIPS, 2023.

ES, S. et al. **Ragas: Automated Evaluation of Retrieval Augmented Generation.** arXiv preprint arXiv:2309.15217, 2023.

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

FOUNDERTOFORTUNE. **Contract-based AI: Moving beyond Prompt Engineering for Reliable Systems.** AI Software Engineering Review, 2025.

JOHNSON, J. et al. **Billion-scale similarity search with GPUs.** IEEE Transactions on Big Data, 2019.

KARRAS, A. et al. **LLMs for Cybersecurity in the Big Data Era: A Comprehensive Review.** MDPI, 2025.

LEVI, M. et al. **Toward Cybersecurity-Expert Small Language Models.** arXiv preprint arXiv:2510.14113, 2025.

LEWIS, P. et al. **Retrieval-augmented generation for knowledge-intensive NLP tasks.** In: NeurIPS, 2020.

MA, L. et al. **AdaptiveLog: An Adaptive Log Analysis Framework with the Collaboration of Large and Small Language Model.** arXiv preprint arXiv:2501.11031, 2025.

MASRI, Y. et al. **Benchmarking Small Language Models and Small Reasoning Language Models on System Log Severity Classification.** arXiv preprint arXiv:2601.07790, 2026.

MDPI. **Domain Knowledge-Infused Synthetic Data Generation for LLM-Based ICS Intrusion Detection: Mitigating Data Scarcity and Imbalance.** Electronics, v. 15, 2026.

MUZAMMIL, S. et al. **Towards Small Language Models for Security Query Generation in SOC Workflows.** arXiv preprint arXiv:2512.06660, 2025.

PALMA, G. et al. **Leveraging Large Language Models for Scalable and Explainable Cybersecurity Log Analysis.** MDPI, 2024.

PALO ALTO NETWORKS RESEARCH. **Why SLMs are Replacing LLMs for Real-time Log Analysis in 2025.** Cyber Defense Insights, 2025.

PRIETO, P. et al. **Edge Deployment of Small Language Models, a comprehensive comparison of CPU, GPU and NPU backends.** arXiv preprint arXiv:2511.22334, 2025.

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

REINSEL, D. et al. **Data Age 2025: The Digitization of the World from Edge to Core.** IDC, 2018.

RODRIGUES, R. F. et al. **Análise de logs de rede sem fio com suporte à tomada de decisão na infraestrutura Tecnologia da informação.** Revista Brasileira de Criminalística, v. 14, 2025.

SAMPAIO, I. P. **Automação de Processos de Resposta a Eventos de Segurança.** 2024.

STROM, B. E. et al. **Mitre att&ck: Design and philosophy.** The MITRE Corporation, 2018.

TALASSO, G. U. et al. **Fine-Tuning Eficiente de Modelos de Linguagem Para Detectar Anomalias em Logs Privados usando Aprendizado Federado.** In: SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES E SISTEMAS DISTRIBUÍDOS (SBRC), 2025.

TANIUM RESEARCH. **The Intent Gap: Challenges in Autonomous AI Security Agents.** Global Cybersecurity Trends, 2026.

TARIQ, M. et al. **Large Language Models in Cybersecurity: A Survey of Applications, Vulnerabilities, and Defense Techniques.** MDPI, 2025.

WANG, Y. et al. **Self-Instruct: Aligning Language Models with Self-Generated Instructions.** In: ACL, 2023.

WEI, B. et al. **CORTEX: Collaborative LLM Agents for High-Stakes Alert Triage.** arXiv preprint arXiv:2510.00311, 2025.

WEI, J. et al. **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.** In: NeurIPS, 2022.

WIZ RESEARCH. **Fine-Tuning a Small Language Model for Secrets Detection in Code.** Wiz, 2025.

WU, J. et al. **When Cache Poisoning Meets LLM Systems: Semantic Cache Poisoning.** In: NDSS Symposium, 2026.

REMUNOM
ISSN 2178-6925

Received: 18/05/2026 - **Accepted:** 21/05/2026
Vol: 13.10
DOI: 10.66104/yv01hn69
Pages: 1-20

YANG, L. et al. **True Attacks, Attack Attempts, or Benign Triggers? Empirical Measurement of Network Alerts in a Security Operations Center.** In: 33rd USENIX Security Symposium, 2024.

ZHANG, J. et al. **When LLMs meet cybersecurity: a systematic literature review.** Cybersecurity, 2025.

ZHENG, L. et al. **Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.** In: Neural Information Processing Systems (NeurIPS), 2023.

ZHONG, A. et al. **LogParser-LLM: Advancing Efficient Log Parsing with Large Language Models.** arXiv preprint, 2024.

ZHOU, J. et al. **Instruction-Following Evaluation for Large Language Models.** arXiv preprint arXiv:2311.07911, 2023.