UFSM

InterAção

ISSN 2357-7975
ACESSO ABERTO

IA., Santa Maria, v. 17, n. 3, e97309, p. 1-34, 2026 • https://dx.doi.org/10.5902/2357797597309
Submissão: 22/06/2026 • Aprovação: 14/09/2026 • Publicação: 02/10/2026

Artigos Dossiê

Ciberdefesa e LLMs no Sul Global: governança, vulnerabilidades e autonomia

Cyberdefense and LLMs in the Global South: governance, vulnerabilities, autonomy

Onildo Ribeiro de Assis II[I]
Ana Isabella Bezerra Lau Ribeiro[I]

[I]Universidade do Tocantins, Paraíso, TO, Brasil

Resumo

Este artigo investiga em que medida a adoção de Modelos de Linguagem de Grande Porte (LLMs) em rotinas de ciberdefesa, especialmente quando consumidos como serviço em nuvem e integrados a fluxos de dados sensíveis, amplia vulnerabilidades e compromete a autonomia estratégica de países do Sul Global. Argumenta-se que a dependência estrutural de LLMs proprietários e de infraestrutura de nuvem centralizada cria três ordens de risco interconectadas: técnico-operacional (vulnerabilidades em sistemas LLM-integrados), jurídico-institucional (questões de jurisdição, governança de dados e auditoria em ambientes não controlados) e geopolítico-estrutural (condicionamento ou negação de acesso em cenários de crise). Utilizando um quadro analítico que combina dependência tecnológica, soberania digital e colonialidade, e artefatos como matrizes de vulnerabilidade, o estudo corrobora a hipótese de que a opacidade, a assimetria de auditoria e a concentração de infraestrutura digital exacerbam esses riscos. Conclui-se que a modernização da ciberdefesa no Sul Global exige que a autonomia seja operacionalizada em requisitos objetivos de governança — como auditabilidade, controle de atualização e acesso a logs — para evitar que a adoção de LLMs se traduza em dependência operacional opaca.

Palavras-chave: LLMs; Ciberdefesa; Autonomia estratégica; Dependência tecnológica, Sul Global

Artigo publicado por InterAção sob uma licença CC BY-NC-SA 4.0.

Abstract

This article investigates the extent to which the adoption of Large Language Models (LLMs) in cyber defense routines, particularly when consumed as a cloud service and integrated into sensitive data flows, amplifies vulnerabilities and compromises the strategic autonomy of Global South countries. It is argued that the structural dependence on proprietary LLMs and centralized cloud infrastructure creates three interconnected orders of risk: technical-operational (vulnerabilities in LLM-integrated systems), legal-institutional (issues of jurisdiction, data governance, and auditing in uncontrolled environments), and geopolitical-structural (conditioning or denial of access in crisis scenarios). Employing an analytical framework that combines technological dependence, digital sovereignty, and coloniality, alongside artifacts such as vulnerability matrices, the study corroborates the hypothesis that opacity, audit asymmetry, and the concentration of digital infrastructure exacerbate these risks. It concludes that the modernization of cyber defense in the Global South requires autonomy to be operationalized into objective governance requirements—such as auditability, update control, and access to logs—to prevent LLM adoption from translating into opaque operational dependence.

Keywords: LLMs; Cyber defense; Strategic autonomy; Technological dependence; Global South

1 INTRODUÇÃO

A vantagem operacional em defesa e segurança tem sido descrita, em documentos doutrinários contemporâneos, como crescentemente dependente de integração de dados, conectividade e comando e controle sob contestação, em vez de derivar apenas do desempenho isolado de plataformas. Essa orientação aparece de forma explícita em formulações norte-americanas associadas a MDO/JADO/JADC2, que vinculam superioridade a ciclos decisórios acelerados e coordenação em rede, e também em doutrinas aliadas que tratam o ciberespaço e o espectro eletromagnético como ambientes permanentes de disputa (ESTADOS UNIDOS, 2018; ESTADOS UNIDOS, 2022; ORGANIZAÇÃO DO TRATADO DO ATLÂNTICO NORTE, 2015; ORGANIZAÇÃO DO TRATADO DO ATLÂNTICO NORTE, 2018). No plano brasileiro, a Doutrina Militar de Defesa descreve dimensões informacionais e cognitivas como transversais a operações em múltiplos domínios, situando o problema da integridade informacional no núcleo de capacidades contemporâneas (BRASIL, 2025).

Nessa mesma moldura, a incorporação de inteligência artificial a rotinas de ciberdefesa passou a ser apresentada, em marcos de gestão de risco, como um campo

que exige governança específica, uma vez que ganhos de automação podem coexistir com novas classes de falhas e com assimetrias de controle, auditoria e transparência. O AI Risk Management Framework do NIST consolida essa preocupação ao tratar risco em IA como fenômeno sociotécnico e ao recomendar práticas de governança, mensuração e gestão contínua ao longo do ciclo de vida (NIST, 2023). A literatura técnica e guias de boas práticas para aplicações com LLMs, por sua vez, explicitam vetores de ataque e falhas recorrentes quando modelos são conectados a ferramentas, bases de conhecimento e fluxos de decisão, em especial injeção de comandos, exposição de dados e riscos de cadeia de suprimentos (OWASP, 2024).

O ponto, aqui, não é sugerir que LLMs sejam intrinsecamente "inadequados" para segurança. A dificuldade começa quando se desloca o olhar do desempenho para a infraestrutura e a governança: em aplicações sensíveis, o que é incorporado não é apenas um artefato algorítmico, mas um arranjo técnico-institucional que inclui modelo, computação, camadas de integração (APIs, plugins, conectores), telemetria e regimes contratuais. Parte desse arranjo pode permanecer opaca ao usuário final, sobretudo quando a solução é consumida como serviço, sem controle efetivo sobre atualizações, treinamento, logs e auditoria independente, o que tensiona recomendações gerais de gestão de risco em IA (NIST, 2023) e amplia superfícies exploráveis descritas em taxonomias recentes de risco para aplicações com LLMs (OWASP, 2024).

Quando essa arquitetura é adotada por Estados do Sul Global, a discussão se aproxima de um vocabulário clássico de dependência tecnológica, com uma diferença: a dependência deixa de ser apenas material-industrial e passa a incluir dependência de modelos e de governança de dados, frequentemente ancoradas em infraestrutura de computação em nuvem controlada por atores externos ao Estado usuário. A noção de *lock-in* — discutida como aprisionamento tecnológico que eleva custos de migração, reduz poder de barganha e limita interoperabilidade — tem sido usada para caracterizar como ecossistemas proprietários podem impor trajetórias tecnológicas e riscos de continuidade (ISSMAEL JÚNIOR, 2026). Essa leitura dialoga com diagnósticos

sobre autonomia tecnológica e soberania em cadeias produtivas de defesa, nos quais a "caixa-preta" de sistemas fechados aparece como restrição a auditoria e à soberania decisória (POCHMANN, 2026), e com análises sobre a Base Industrial de Defesa brasileira que enfatizam dependência externa, fragilidade de cadeias e necessidade de coordenação estatal de longo prazo (NEIVA FILHO; SILVA, 2026).

A inserção do Brasil nesse debate não é meramente especulativa. A Estratégia Nacional de Cibersegurança, instituída por decreto, estabelece diretrizes voltadas à resiliência de serviços essenciais e infraestruturas críticas, apontando para a centralidade do controle e da continuidade em ambientes contestados (BRASIL, 2025a). Em paralelo, a Política Nacional de Defesa, a Estratégia Nacional de Defesa e o Livro Branco, aprovados também por decreto, reafirmam a autonomia tecnológica como condicionante da capacidade de defesa e como eixo de soberania, ainda que a problemática específica de dependência em *foundation models* e em infraestrutura de nuvem para IA não apareça com o mesmo grau de detalhamento (BRASIL, 2025b). O resultado é um descompasso plausível: incentiva-se modernização e digitalização, mas o debate público sobre requisitos de auditabilidade, soberania de dados e continuidade operacional para IA em funções críticas tende a permanecer fragmentado.

A camada técnica ajuda a explicar por que essa fragmentação custa caro. Classes de vulnerabilidade já descritas na literatura de segurança em aprendizado de máquina incluem ataques adversariais, contaminação de dados e exploração de falhas sistêmicas que, em contextos de decisão, podem produzir resultados errados com aparência de normalidade; o panorama é sintetizado em revisões que tratam a segurança de ML como campo no qual ataque e defesa coevoluem (BIGGIO; ROLI, 2018). Em aplicações com LLMs conectadas a ferramentas e dados externos, a injeção de comandos (prompt injection) ganhou destaque por explorar o fato de que texto pode carregar instruções adversariais capazes de subverter comportamento do sistema, inclusive quando o modelo atua como "orquestrador" de ferramentas; esse vetor foi sistematizado em trabalho que discute a injeção como ameaça relevante a

aplicações LLM-integradas (GRESHake et al., 2023). Já o vazamento de dados sensíveis, por memorização e extração, foi demonstrado empiricamente em pesquisa sobre extração de dados de treinamento a partir de grandes modelos de linguagem, reforçando que confidencialidade pode ser degradada sem que o usuário perceba a tempo (CARLINI et al., 2021). Nada disso, por si, "condena" o uso de IA; mas sugere que, sem governança e controles, o risco não se limita ao erro pontual: o risco passa a ser sistêmico e acumulativo, como reconhecem guias de gestão de risco e taxonomias voltadas a aplicações reais (NIST, 2023; OWASP, 2024).

Essa constelação de fatores conduz ao problema de pesquisa. Se capacidades de ciberdefesa são fortalecidas por tecnologia, a dependência de ecossistemas fechados e infraestruturas externas pode reintroduzir vulnerabilidades por vias pouco intuitivas, inclusive em funções de triagem, análise e apoio à decisão. Em termos mais diretos, um Estado pode comprar velocidade e perder controle. A hipótese adotada é que a dependência estrutural de LLMs proprietários e de infraestrutura de nuvem centralizada cria três ordens de risco que se reforçam mutuamente: (i) risco técnico, associado a vulnerabilidades e ataques em sistemas LLM-integrados; (ii) risco jurídico-operacional, relacionado a jurisdição, governança de dados, auditoria e resposta a incidentes em ambientes não plenamente controlados; (iii) risco geopolítico, vinculado à possibilidade de condicionamento, negação de acesso ou coerção indireta em cenários de crise, dinâmica compatível com leituras de *lock-in* e dependência discutidas no próprio dossiê (ISSMAEL JÚNIOR, 2026; NEIVA FILHO; SILVA, 2026; POCHMANN, 2026).

A pergunta de pesquisa é formulada nestes termos: em que medida a adoção de LLMs baseados em infraestrutura centralizada do Norte Global expõe o Brasil a vulnerabilidades estratégicas no domínio da ciberdefesa? O objetivo geral consiste em mapear vulnerabilidades técnicas relevantes para o emprego de LLMs em rotinas de ciberdefesa e conectá-las a um quadro analítico de dependência tecnológica e soberania digital, de modo compatível com diretrizes nacionais de autonomia e

resiliência (BRASIL, 2025a; BRASIL, 2025b). Três objetivos específicos organizam essa tarefa: primeiro, sistematizar usos e modalidades de adoção de LLMs em ciberdefesa à luz de marcos de risco em IA (NIST, 2023; OWASP, 2024); segundo, consolidar classes de vulnerabilidade tecnicamente descritas na literatura, com ênfase em prompt injection, vazamento e contaminação de dados, bem como riscos de cadeia de suprimentos (BIGGIO; ROLI, 2018; GRESHake et al., 2023; CARLINI et al., 2021); terceiro, articular essas classes a uma leitura estratégica inspirada em debates recentes do campo de defesa no Sul Global sobre autonomia, governança e dependência em ecossistemas tecnológicos (ISSMAEL JÚNIOR, 2026; NEIVA FILHO; SILVA, 2026; POCHMANN, 2026).

O artigo está organizado em quatro movimentos. Inicialmente, descrevem-se usos de LLMs em ciberdefesa e as condições típicas de adoção em contextos de restrição fiscal e heterogeneidade de sistemas, cenário frequentemente associado, no debate do dossiê, a soluções proprietárias e trajetórias de dependência (ISSMAEL JÚNIOR, 2026; NEIVA FILHO; SILVA, 2026). Em seguida, apresenta-se o referencial teórico que combina dependência tecnológica, soberania digital e colonialidade, aproximando-o de problemas de governança e opacidade (“caixa-preta”) observáveis em ecossistemas fechados (POCHMANN, 2026; NIST, 2023). A metodologia é explicitada como revisão bibliográfica e documental, com síntese em matrizes analíticas (NIST, 2023; OWASP, 2024). Por fim, os resultados e a discussão conectam classes técnicas de vulnerabilidade a perdas de verificabilidade (logs, auditoria, controle de versão e continuidade), articulando implicações jurídico-operacionais e geopolíticas para a autonomia em ciberdefesa.

2 DESENVOLVIMENTO

LLMs vêm sendo acoplados a rotinas de segurança porque reduzem fricções de linguagem entre sinais técnicos e decisões organizacionais: um evento de SIEM pode ser reescrito como narrativa operacional; um relatório de inteligência pode ser sumarizado sob restrições de tempo; um analista pode iterar hipóteses em linguagem

natural. A utilidade, porém, não se distribui no vácuo. Quando o modelo é consumido como serviço — com telemetria, atualizações e dependências fora do alcance do usuário —, o ganho cognitivo local passa a carregar custos sistêmicos de auditoria e continuidade, o que desloca a discussão de "eficiência" para governança e soberania, tema recorrente no dossiê ao tratar de aprisionamento tecnológico e resiliência sob contestação (ISSMAEL JÚNIOR, 2026).

A literatura recente de gestão de risco em IA insiste nesse ponto: risco não é apenas falha técnica; ele é sociotécnico, deriva de ciclo de vida, cadeias de suprimentos, arranjos de responsabilidade e métricas de controle. Ao sistematizar funções de governança e avaliação, o AI RMF do NIST dá linguagem para tratar o problema sem recorrer a suposições heroicas de confiabilidade do modelo (NIST, 2023). Em paralelo, taxonomias de risco orientadas a aplicações com LLMs já estabilizaram um vocabulário mínimo de vulnerabilidades recorrentes quando modelos são integrados a ferramentas, dados e processos organizacionais, com ênfase em injeção de comandos, exposição de informações e fragilidades de integração (OWASP, 2024).

O caso brasileiro entra por duas portas. A primeira é normativa: há diretrizes estatais que colocam resiliência e proteção de serviços essenciais e infraestruturas críticas como prioridades, criando um critério de avaliação que não se satisfaz com "bons resultados médios" (BRASIL, 2025a). A segunda é estrutural: uma agenda de autonomia tecnológica em defesa é reiterada como condição de soberania, ainda que o debate público sobre dependência em modelos fundacionais e infraestrutura de nuvem para IA permaneça menos detalhado do que outras dimensões de capacidade (BRASIL, 2025b). A Base Industrial de Defesa, analisada como ecossistema heterogêneo e com desafios de dependência externa, oferece o pano de fundo material: capacidades existem, mas convivem com gargalos de cadeia produtiva e previsibilidade de políticas, o que condiciona qualquer estratégia de "internalizar" tecnologias críticas (NEIVA FILHO; SILVA, 2026).

Este desenvolvimento, portanto, não trata LLM como peça isolada. O que será descrito é um arranjo técnico-institucional típico de adoção, e como esse arranjo reorganiza vulnerabilidades na ciberdefesa: parte delas nasce em classes de ataque já discutidas em segurança de ML; outra parte é produzida pela própria arquitetura de integração e por regimes de governança e jurisdição (BIGGIO; ROLI, 2018; NIST, 2023; OWASP, 2024).

2.1 LLMs na ciberdefesa: usos, potencialidades e adoção no Sul Global

O emprego de modelos de linguagem de grande porte (LLMs) em ciberdefesa tem sido proposto como apoio à decisão e como mecanismo de redução de fricção cognitiva em tarefas intensivas em texto: triagem de alertas, consolidação de evidências, redação de relatórios e tradução de achados técnicos para linguagem operacional. Na prática, essa incorporação tende a ocorrer como camada adicional sobre rotinas já institucionalizadas — gestão de logs, tratamento de incidentes e gestão de vulnerabilidades —, em vez de substituí-las integralmente (NIST, 2006; NIST, 2012; NIST, 2013). Sob a lente de gestão de risco, o ponto sensível não é apenas "o modelo pode errar", mas o fato de que o sistema passa a depender de ciclo de vida, governança e integrações (NIST, 2023).

Em SOC/CSIRT, a utilidade mais defensável do LLM aparece quando ele é mantido como instrumento de normalização e síntese, e não como árbitro final do evento. A triagem e o registro consistente de incidentes são apresentados pelo NIST como requisitos de coordenação e rastreabilidade, e não como formalismo; daí por que a saída do LLM precisa ser tratada como rascunho verificável, com validação humana e evidência preservada (NIST, 2012). De modo semelhante, na análise e correlação de logs, o NIST enfatiza integridade, retenção, sincronização temporal e revisão contínua como base de auditoria e resposta; por isso, o uso de LLMs tende a ser mais seguro quando se restringe a explicar campos, apoiar a formulação de consultas e sumarizar resultados já extraídos por mecanismos determinísticos de log management (NIST,

2006). Na gestão de vulnerabilidades, a função típica é auxiliar triagem de advisories e priorização inicial, mas o próprio NIST trata patch management como problema de governança, inventário e testes, o que limita o espaço para automatismos não auditáveis (NIST, 2013).

O quadro muda quando o LLM é acoplado a ferramentas e repositórios (por exemplo, conectores com SIEM/SOAR, bases internas via RAG, sistemas de ticket e rotinas automatizadas). Nesse cenário, o risco passa a se concentrar na superfície de integração, que já foi sistematizada em taxonomias voltadas a aplicações com LLMs, com ênfase em prompt injection, exposição de dados sensíveis e fragilidades de controle de acesso e de ferramentas (OWASP, 2024). A injeção indireta de prompt — em que instruções adversariais são inseridas em conteúdos "lidos" pelo sistema e acabam subvertendo o comportamento do LLM integrado — foi descrita e demonstrada como classe de ameaça relevante a aplicações reais LLM-integradas (GRESHake et al., 2023). Em paralelo, riscos de confidencialidade não são meramente hipotéticos: há demonstração empírica de extração de dados de treinamento em grandes modelos de linguagem, o que reforça que minimização e governança de dados sensíveis são requisitos em contextos críticos (CARLINI et al., 2021).

A adoção no Sul Global pode ser interpretada como uma escolha entre arranjos com diferentes graus de governança efetiva. Quando consumido como serviço em nuvem, o LLM tende a oferecer rápida escalabilidade, mas frequentemente desloca para fora do perímetro organizacional elementos essenciais de controle: telemetria, políticas de retenção, atualização, e parte da auditabilidade prática do sistema, exatamente as dimensões que o AI RMF destaca como centrais para gestão de risco ao longo do ciclo de vida (NIST, 2023). Em configurações on-premise (ou em nuvem privada sob controle estatal/organizacional), aumenta a margem para auditoria e continuidade, mas o ônus operacional recai sobre a instituição, que precisa sustentar atualização, segurança e resposta a incidentes com capacidade própria (NIST, 2012; NIST, 2023). Em modelos open-source com execução e ajuste locais, a possibilidade

de adaptação e maior visibilidade do artefato é ampliada; a disponibilização de modelos com pesos públicos é exemplificada pelo Llama 2 (TOUVRON et al., 2023). Ainda assim, a abertura do modelo não elimina riscos de aplicação: vetores como injeção de prompt e falhas de integração persistem, pois derivam do arranjo sistema–dados–ferramentas, não apenas do regime de licenciamento (GRESHake et al., 2023; OWASP, 2024).

Por fim, há uma tensão estrutural que ajuda a explicar por que a dependência reaparece mesmo quando se reconhece o risco: modelos de alto desempenho exigem computação intensiva e ecossistemas maduros de ferramentas, o que favorece "pacotes" integrados. Esse padrão se insere em dinâmicas mais amplas da economia digital, nas quais concentração de infraestrutura e assimetrias de governança de dados são discutidas como características com implicações diretas para países em desenvolvimento (UNCTAD, 2021). Em ciberdefesa, a consequência é direta: a mesma arquitetura que acelera ganhos locais pode ampliar dependências difíceis de reverter, justamente em um domínio em que resiliência e continuidade sob contestação são requisitos, não preferências (NIST, 2010; NIST, 2023).

2.2 Fundamentos teóricos: dependência tecnológica, soberania digital e colonialidade

O problema analisado não é "adoção de IA" em abstrato, mas a incorporação de capacidades cognitivas automatizadas em uma pilha sociotécnica frequentemente externa (modelo, computação, nuvem, integrações e regime de atualização). Em ciberdefesa, esse arranjo importa porque desloca o centro de gravidade do risco: deixa de ser apenas uma questão de desempenho do artefato e passa a envolver governança do ciclo de vida, rastreabilidade e capacidade de auditoria. O AI Risk Management Framework do NIST é particularmente útil aqui por definir risco em IA como fenômeno sociotécnico e por enfatizar governança contínua, monitoramento e gestão ao longo do ciclo de vida (NIST, 2023). Assim, um Estado pode "usar" um

sistema e, ainda assim, deter controle incompleto sobre como ele muda, como falha e o que registra, sobretudo quando a solução é consumida como serviço e permanece parcialmente opaca ao usuário (NIST, 2023).

A tradição da dependência permite tratar essa assimetria como relação estrutural, e não como acidente. O esquema centro–periferia, em Prebisch, já vinculava desenvolvimento desigual à distribuição internacional de capacidade produtiva e tecnológica, com efeitos persistentes sobre poder de decisão e trajetórias de modernização (PREBISCH, 1950). Em Cardoso e Faletto, dependência descreve a combinação entre condicionantes externos e arranjos internos que preservam heteronomia e limites de decisão, mesmo na presença de modernização (CARDOSO; FALETTO, 1979). Em Dos Santos, a dependência é formulada como vínculo no qual a expansão e as opções de uma economia ficam condicionadas à dinâmica de outra (DOS SANTOS, 1970). Transposta para LLMs em ciberdefesa, essa linguagem não sugere "incapacidade total", mas um ganho funcional que pode coexistir com heteronomia de governança: o núcleo cognitivo e parte do regime de atualização e observabilidade permanecem fora de alcance pleno.

A noção de soberania digital entra, então, como problema prático de jurisdição e controle do ciclo informacional (dados, processamento, armazenamento e logs). Em contextos estatais, a soberania de dados não é apenas um ideal: ela se conecta a obrigações de proteção e a requisitos de governança. No caso brasileiro, a LGPD estabelece princípios e bases legais para tratamento de dados pessoais, impondo deveres e limites que se tornam relevantes quando fluxos de dados atravessam provedores e fronteiras (BRASIL, 2018). Em paralelo, diretrizes nacionais recentes reiteram resiliência e autonomia tecnológica como objetivos estruturantes no campo de defesa e cibersegurança, criando um critério político para avaliar terceirizações profundas em capacidades críticas (BRASIL, 2025a; BRASIL, 2025b). Do lado externo, dispositivos como o CLOUD Act explicitam tensões de jurisdição sobre dados sob controle de provedores sujeitos à lei norte-americana, ainda que armazenados fora

do território, elemento frequentemente discutido em debates sobre governança de nuvem e accountability (UNITED STATES, 2018).

A terceira peça do referencial é a colonialidade tecnológica, útil para nomear um tipo específico de assimetria: não apenas "comprar" tecnologia, mas operar sob um regime em que critérios de validação e capacidade de verificação independente são limitados. A colonialidade, em Quijano, descreve padrões duráveis de hierarquia e de produção de saber/poder que ultrapassam a dimensão estritamente econômica (QUIJANO, 2000). No plano do digital, Couldry e Mejias argumentam que a extração e apropriação de dados configuram uma forma contemporânea de colonialismo, em que a infraestrutura de captura e processamento tende a concentrar poder (COULDRY; MEJIAS, 2019). Essa discussão se articula diretamente com o tema da caixa-preta: Pasquale descreve como decisões mediadas por sistemas opacos dificultam accountability e contestação (PASQUALE, 2015), e Burrell sistematiza formas de opacidade por segredo, por complexidade e por déficit de letramento técnico (BURRELL, 2016). Em LLMs, esse pacote é operacional: o usuário recebe saídas linguísticas persuasivas, mas pode não controlar treinamento, ajustes, telemetria e atualização — e, em contextos de investigação e resposta, reprodutibilidade e rastreabilidade não são luxos (NIST, 2023).

Há, por fim, um eixo de mitigação que reforça o argumento teórico: documentação como redução de assimetria. Propostas como *datasheets for datasets* e *model cards* foram formuladas para padronizar informações sobre conjuntos de dados e modelos, visando permitir avaliação mais responsável e reduzir opacidades evitáveis (GEBRU et al., 2018; MITCHELL et al., 2019). O ponto não é supor que documentação elimine risco; é reconhecer que, sem documentação e sem direitos de auditoria, a periferia tende a operar por confiança em autoridade e por garantias contratuais mínimas — uma forma prática de colonialidade como assimetria de verificação (BURRELL, 2016; PASQUALE, 2015).

Com isso, o referencial é convertido em lente analítica para as seções seguintes. Dependência será tratada como limitação de auditabilidade, lock-in e custos de saída — no sentido clássico de aprisionamento e elevação de custos de migração discutido na literatura econômica (FARRELL; KLEMPERER, 2007). Soberania digital será operacionalizada como jurisdição aplicável, localização e controle de processamento/armazenamento e acesso a logs, em coerência com obrigações e diretrizes nacionais (BRASIL, 2018; BRASIL, 2025a; BRASIL, 2025b). Colonialidade será tomada como caixa-preta, assimetria epistêmica e impossibilidade de verificação independente, com deslocamento de accountability para fora do controle estatal, ainda que a responsabilidade permaneça interna (QUIJANO, 2000; COULDRY; MEJIAS, 2019; PASQUALE, 2015).

2.3 Vulnerabilidades técnicas de LLMs em ciberdefesa e infraestrutura crítica

A discussão de vulnerabilidades em LLMs, quando aplicada à ciberdefesa e à proteção de infraestrutura crítica, não se reduz a "riscos gerais de IA". O ponto decisivo é arquitetural: LLMs tendem a operar acoplados a dados operacionais sensíveis e a camadas de automação e integração (por exemplo, conectores, ferramentas e bases de conhecimento), o que altera a superfície de ataque e as condições de controle. Nessa configuração, risco não decorre apenas do conteúdo gerado, mas do ciclo de vida do sistema e da forma como ele é governado e monitorado (NIST, 2023). Por isso, a leitura "desempenho médio alto ⇒ risco baixo" é insuficiente: governança, monitoramento, métricas e responsabilidade permanecem centrais (NIST, 2023).

Uma síntese útil das classes de falha em aplicações com LLMs já foi estabilizada em taxonomias técnicas voltadas a uso real, destacando vetores como injeção de prompt, vazamento de dados, excesso de permissões em ferramentas, e riscos de dependências/integrações (OWASP, 2024). Em ciberdefesa, essas categorias funcionam como mapa de ataque precisamente porque o LLM pode se tornar mediador de triagem, priorização e linguagem operacional — e, em sistemas críticos, a mediação tem efeitos materiais.

Entre esses vetores, o envenenamento de dados (data poisoning) permanece relevante como classe geral de ataque a aprendizado de máquina: trata-se da inserção deliberada (ou contaminação) de dados que degrada desempenho ou induz comportamentos direcionados. A literatura descreve ataques de poisoning tanto de impacto global (degradação geral) quanto seletivo (pontos cegos), com trade-offs próprios (BIGGIO; NELSON; LASKOV, 2012; STEINHARDT; KOH; LIANG, 2017). Para ciberdefesa, a seletividade é a forma operacionalmente mais perigosa: não é necessário "derrubar" o processo inteiro; basta induzir omissão ou distorção em padrões específicos de interesse do adversário (BIGGIO; NELSON; LASKOV, 2012; STEINHARDT; KOH; LIANG, 2017). Em sistemas LLM-integrados, essa lógica se amplia porque a contaminação não se limita ao treino original: pode ocorrer em etapas de ajuste/curadoria e, de forma particularmente plausível, nas fontes externas que alimentam o sistema (por exemplo, bases e documentos consultados), onde o atacante tenta manipular o que entra como "contexto" para influenciar a saída (OWASP, 2024; GRESHAKE et al., 2023).

A injeção de prompt, especialmente em sua forma indireta (quando instruções adversariais são inseridas em conteúdos consultados pelo sistema), foi discutida como ameaça relevante a aplicações LLM-integradas porque explora o fato de que o modelo pode tratar texto como instrução e subverter o comportamento esperado, inclusive em cenários com ferramentas e recuperação de informação (GRESHAKE et al., 2023). Em contexto de ciberdefesa, o impacto esperado não é "conteúdo inadequado" em sentido moral, mas recomendação operacionalmente perigosa: omitir evidências, reduzir severidade, sugerir ação de contenção imprópria ou orientar investigações para trilhas falsas, o que é compatível com a ênfase da OWASP em riscos decorrentes de integração e de excesso de confiança no componente linguístico (OWASP, 2024).

O vazamento de dados também tem base empírica na literatura: há demonstrações de extração de dados de treinamento a partir de grandes modelos de linguagem, reforçando que confidencialidade pode ser degradada por memorizações

e mecanismos de extração, o que é crítico quando o sistema processa informações sensíveis (CARLINI et al., 2021). Em aplicações operacionais, o risco se materializa tanto por exposição direta (inputs/outputs) quanto por combinações de integração e permissões que ampliam o que o modelo pode acessar e revelar (OWASP, 2024).

Essas vulnerabilidades técnicas se conectam diretamente ao tema de dependência estrutural porque seu impacto é condicionado pelo modo de adoção. Quando o LLM é consumido como serviço proprietário e remoto, quatro capacidades tendem a ficar reduzidas para o usuário — capacidades que são centrais para segurança em ambientes críticos: (i) visibilidade, pois logs e telemetria podem ser limitados ou insuficientes para reconstruir eventos; (ii) auditabilidade, pois dados de treinamento/ajuste e decisões de engenharia permanecem opacos; (iii) controle de atualização, pois mudanças de versão podem alterar comportamento sem possibilidade simples de congelamento, rollback ou reprodutibilidade operacional; e (iv) red teaming pleno, pois testes podem ser limitados por termos de uso, ausência de acesso às camadas internas ou restrições contratuais (NIST, 2023; BURRELL, 2016). Burrell é útil para caracterizar a opacidade como fenômeno que pode resultar de segredo corporativo, complexidade técnica e barreiras de compreensão — combinação que tende a se intensificar quando o sistema é ofertado como serviço (BURRELL, 2016). O AI RMF do NIST, ao enfatizar governança contínua e gestão do ciclo de vida, fornece o enquadramento institucional para tratar essas perdas como risco, e não como “detalhe de implementação” (NIST, 2023).

2.4 Governança e accountability de LLMs em ciberdefesa no Sul Global

Em ciberdefesa — e, com mais razão, em infraestrutura crítica — o problema não é apenas “segurança do modelo”, mas governança do arranjo: quem responde por falhas, quem audita, quem pode congelar versões, quem acessa logs e evidências, o que ocorre quando o serviço degrada ou cai e quais dados sensíveis atravessam fronteiras. No Sul Global, essa governança é pressionada por restrições materiais (compute, pessoal, financiamento) e por assimetrias contratuais e de poder de

mercado, que reduzem a capacidade de impor exigências sem perder acesso a ferramentas de alto desempenho.

O AI Risk Management Framework do NIST fornece um ponto de apoio direto para essa discussão ao caracterizar risco em IA como sociotécnico e ao tratar governança como função contínua, vinculada a mensuração, monitoramento e gestão do ciclo de vida — e não como etapa inicial "documental" (NIST, 2023). Isso se ajusta ao cotidiano de SOC/CSIRT: incidentes mudam, integrações envelhecem, conectores quebram, e versões alteram comportamento. Se a governança não for operacionalizável e testável, ela não se sustenta.

A primeira fricção é a responsabilidade. Recomendações de um LLM (priorização, hipótese de causa raiz, sugestão de contenção) podem influenciar tempo de resposta e alocação de recursos. Quando uma recomendação é equivocada e produz dano, a responsabilização tende a recair sobre a organização usuária, enquanto o fornecedor se posiciona como provedor de ferramenta. Em IA, isso é agravado por opacidade e variação entre versões. Por isso, accountability, aqui, significa capacidade de demonstrar como uma saída foi produzida: com que dado, por qual versão, em qual configuração e com quais integrações ativas (NIST, 2023). Sem essa rastreabilidade, o pós-incidente vira disputa narrativa.

Daí decorre um requisito mínimo: trilha de auditoria e evidências. Práticas clássicas de gestão de logs e resposta a incidentes permanecem centrais porque permitem reconstruir eventos, testar hipóteses e aprender institucionalmente (NIST, 2006; NIST, 2012). Em aplicações com LLMs, isso precisa incluir, além de telemetria usual, registros de prompts, contexto recuperado (quando houver RAG), saídas do modelo e ações executadas por ferramentas, com retenção e controles compatíveis com o nível de criticidade (NIST, 2006; NIST, 2023).

O segundo requisito é controle de acesso e privilégio mínimo nas integrações. Em sistemas "tool-augmented", o risco frequentemente migra do texto para as permissões: o modelo não precisa "invadir" nada; basta induzir o sistema a acionar

uma ferramenta com credenciais excessivas. Taxonomias de risco para aplicações com LLMs tratam precisamente de excesso de permissões, falhas de integração e exposição de dados como classes recorrentes (OWASP, 2024). Em termos práticos, governança significa segmentação, escopo restrito, aprovação humana para ações sensíveis e separação entre ambientes de análise e de execução.

O terceiro requisito é validação de entradas e higiene de fontes. Como injeção direta e indireta de prompt é um vetor estabilizado na literatura técnica aplicada a LLM apps, entradas oriundas de fontes não confiáveis (tickets, e-mails, páginas, repositórios, artefatos coletados) devem ser tratadas como potencialmente hostis, mesmo quando "parecem" dados operacionais (OWASP, 2024). Isso implica políticas formais sobre: quais fontes podem alimentar o modelo; como conteúdos são filtrados/higienizados; e como o sistema reage a padrões de tentativa de subversão.

O quarto requisito é testes contínuos e red teaming. A gestão de risco em IA, no AI RMF, é definida como processo contínuo, com mensuração e monitoramento ao longo do ciclo de vida (NIST, 2023). Para LLMs em ciberdefesa, isso se traduz em rotinas periódicas de teste de prompt injection, avaliação de vazamento de contexto, testes de robustez frente a entradas adversariais e validação de integrações com ferramentas, especialmente após mudanças (modelo, embeddings, corpus RAG, conectores). Se o sistema muda, o risco muda junto — e pode mudar sem alarde (NIST, 2023).

Por fim, há a camada de requisitos contratuais e de jurisdição, que transforma governança em obrigação verificável. Quando a capacidade é contratada externamente, devem existir condições claras sobre localização e retenção de dados, acesso a logs, notificação de incidentes, mecanismos de congelamento/rollback de versão em períodos críticos, parâmetros de manutenção, portabilidade e continuidade. No caso brasileiro, a Estratégia Nacional de Cibersegurança, ao enfatizar resiliência e proteção de serviços essenciais e infraestruturas críticas, oferece base normativa para tratar tais requisitos como condição de contratação em contextos sensíveis (BRASIL, 2025a).

O resultado desses elementos não é "eliminar" risco, mas impedir que a adoção

de LLMs converta uma capacidade de ciberdefesa em dependência operacional opaca. Em ambientes do Sul Global, onde restrições são reais, isso marca a diferença entre modernização governada (com autonomia verificável) e modernização por terceirização de funções críticas (NIST, 2023; OWASP, 2024).

3 METODOLOGIA

Esta pesquisa é qualitativa, analítica e propositiva. O objetivo metodológico não foi testar desempenho de um modelo específico nem produzir evidência experimental por meio de execução de ataques. A opção foi outra: organizar, a partir de fontes abertas e verificáveis, um quadro de riscos e implicações de soberania aplicável ao emprego de LLMs em ciberdefesa e infraestrutura crítica, com ênfase naquilo que pode ser auditado, governado e contratado.

O procedimento central combinou revisão bibliográfica e análise documental. A análise documental foi adotada porque o problema em discussão é fortemente mediado por textos normativos, estratégicos e padrões técnicos, isto é, por documentos que definem obrigações, responsabilidades e limites de governança. Nessa abordagem, documentos não são tratados como “contexto”, mas como dados: o que afirmam, o que exigem, o que deixam em aberto e quais capacidades pressupõem (BOWEN, 2009). A revisão bibliográfica, por sua vez, foi utilizada para identificar classes de ataque e modos de falha associados a sistemas com LLMs e, em paralelo, para sustentar a ponte conceitual com dependência tecnológica, soberania digital e colonialidade tecnológica.

A seleção de fontes foi orientada por três critérios de inclusão. Primeiro, priorizaram-se publicações técnicas e científicas que descrevem mecanismos de vulnerabilidade e/ou taxonomias de risco para aplicações com LLMs, com rastreabilidade (periódicos, anais reconhecidos, preprints com circulação consolidada e documentos de comunidades técnicas). Nesse bloco, foram mobilizadas como âncoras de organização do risco a taxonomia da OWASP para aplicações com LLMs, por

sua função de mapeamento de vetores recorrentes em sistemas integrados (OWASP, 2024), e o AI Risk Management Framework do NIST, por estabelecer linguagem de governança, mensuração e monitoramento ao longo do ciclo de vida de sistemas de IA (NIST, 2023). Segundo, foram incluídos documentos brasileiros de estratégia e normatização que incidem sobre autonomia tecnológica, resiliência de serviços essenciais/infraestruturas críticas e governança de dados, com destaque para a Estratégia Nacional de Cibersegurança (BRASIL, 2025a), o decreto que aprova PND/END/Livro Branco (BRASIL, 2025b) e a LGPD quando pertinente à discussão de tratamento e circulação de dados (BRASIL, 2018). Terceiro, foram incorporadas referências teóricas clássicas e contemporâneas sobre dependência e assimetria tecnológica, utilizadas aqui como ferramentas de interpretação estrutural — não como ornamento.

A extração e o tratamento do material seguiram um roteiro de codificação temática, com categorias definidas de forma iterativa. Num primeiro momento, as categorias foram deduzidas do problema de pesquisa (vulnerabilidades técnicas; arranjos de implantação; governança; jurisdição; auditabilidade; continuidade). Em seguida, elas foram refinadas à medida que a literatura técnica e os documentos estratégicos eram lidos, para reduzir redundâncias e forçar precisão conceitual. A análise de conteúdo foi empregada nesse ponto como técnica de sistematização: trechos foram associados a categorias, tensões e implicações, mantendo-se o cuidado de não deslocar o sentido original das fontes (BARDIN, 2011). O produto desse processo não é um inventário exaustivo de todos os riscos possíveis, mas uma tipologia suficientemente estável para sustentar proposições de governança.

As limitações do estudo são diretas. Não foram utilizados dados operacionais classificados nem se pretendeu inferir, com base em fontes abertas, o grau real de adoção de LLMs em órgãos militares e civis brasileiros. Além disso, o recorte em fontes abertas implica que certos aspectos cruciais — arquitetura interna de serviços proprietários, telemetria efetivamente coletada, condições contratuais específicas e procedimentos internos de resposta a incidentes — não podem ser avaliados com

completude. Por fim, como não foram executados testes ofensivos ou exercícios de red teaming em sistemas reais, os resultados devem ser lidos como mapeamento conceitual fundamentado e como proposta de governança, não como comprovação empírica de exploração em um ambiente específico.

4 RESULTADOS E DISCUSSÃO

Os resultados são apresentados como artefatos de síntese (Quadros 1, 2 e 3) e, em seguida, interpretados à luz do referencial de dependência tecnológica, soberania digital e colonialidade (seção 2.2). O pressuposto analítico é explícito: o risco não reside apenas no "erro" do modelo, mas no conjunto modelo–dados–integrações–ciclo de atualização–governança, conforme a abordagem sociotécnica de risco em IA proposta no AI RMF do NIST (NIST, 2023). As classes de ataque e modos de falha foram organizados com base em taxonomias e evidências abertas sobre aplicações com LLMs (OWASP, 2024; GRESHAKE et al., 2023), bem como na literatura de ataques a aprendizado de máquina (BIGGIO; NELSON; LASKOV, 2012; STEINHARDT; KOH; LIANG, 2017; SHOKRI et al., 2017; CARLINI et al., 2021; TRAMÈR et al., 2016).

A coluna "risco relativo em nuvem remota vs infraestrutura soberana" não expressa prevalência empírica; é um juízo de plausibilidade operacional derivado de quatro fatores que alteram a superfície de ataque e a capacidade de resposta: visibilidade (logs/telemetria), auditabilidade (dados de treino/ajuste), controle de atualização e possibilidade de red teaming pleno. Esses fatores são consistentes com boas práticas de gestão de logs, tratamento de incidentes e continuidade (NIST, 2006; NIST, 2012; NIST, 2010), e com a ênfase do AI RMF em governança e monitoramento contínuos (NIST, 2023).

O quadro 1 sintetiza os achados da revisão documental em uma matriz que relaciona (i) classes de vulnerabilidade em aplicações com LLMs, (ii) efeitos operacionais plausíveis em rotinas de SOC/CSIRT e em contextos de infraestrutura crítica e (iii) pontos de governança que determinam se o risco permanece gerenciável ou se se

converte em dependência operacional. A seleção e a nomenclatura das classes partem de taxonomias abertas voltadas a aplicações com LLMs, especialmente no que se refere a injeção de prompt, vazamento de dados, dependências e excesso de permissões em integrações (OWASP, 2024). A leitura dos impactos e das respostas prioriza requisitos de gestão de risco e ciclo de vida — governança, monitoramento, métricas e responsabilidade — conforme o AI RMF (NIST, 2023), além de práticas de resposta a incidentes, gestão de logs e continuidade, relevantes quando a capacidade analítica passa a depender de sistemas atualizáveis e integrados (NIST, 2006; NIST, 2012; NIST, 2010).

Em termos analíticos, a matriz explicita o argumento central do artigo: em arquiteturas de "caixa-preta", vulnerabilidades conhecidas tendem a ter maior impacto e menor verificabilidade, porque visibilidade (logs/telemetria), auditabilidade e controle de atualização não são plenamente controlados pelo operador.

Quadro 1 – Matriz de vulnerabilidades (LLMs em ciberdefesa/infra crítica)

Classe de ataque	Mecanismo de exploração	Impacto potencial	Dificuldade de detecção	Risco relativo	Mitigações
Injeção de prompt (direta)	Entrada maliciosa subverte instruções, revelando informações ou gerando ações indevidas.	Desvio de triagem, erros de contenção, exposição de dados.	Alta (parece legítima).	Maior em nuvem remota (menor observabilidade).	Validação de entradas, monitoramento de prompts/ saídas, políticas de conteúdo.
Injeção de prompt indireta (via conteúdo recuperado/RAG)	Conteúdo externo (web, tickets, etc.) carrega instruções ocultas que o LLM "executa".	Contaminação decisória, extração de segredos, manipulação de recomendações.	Muito alta (instrução embutida em dado "normal").	Maior em nuvem remota (cadeia de ingestão menos auditável).	Higienização de corpus, isolamento de ferramentas, "policy layer".
Envenenamento de dados (poisoning) – treino/ajuste	Inserção de exemplos maliciosos no treino/ajuste para induzir comportamento desejado.	"Pontos cegos" em detecção, viés sistemático, degradação silenciosa.	Alta (efeito gradual).	Maior em nuvem remota (usuário não audita dados de ajuste).	Governança de dados, curadoria, testes de regressão, controle de versões.

Continuação ...

Quadro 1 – continuação

Classe de ataque	Mecanismo de exploração	Impacto potencial	Dificuldade de detecção	Risco relativo	Mitigações
Envenenamento de dados (poisoning) – base de conhecimento/RAG	Contaminação de repositórios consultados (wikis, runbooks) para induzir respostas/ ações.	Recomendação de *playbooks* errados, normalização de TTPs adversárias.	Alta.	Maior em nuvem remota (gestão do corpus terceirizada/menos observável).	Controle de alterações, assinaturas/ hashes, revisão por pares, *logging* de acesso.
Vazamento de dados / exposição de informações sensíveis	*Prompting*, falhas de isolamento, *logs* indevidos, respostas com dados confidenciais.	Exfiltração indireta, quebra de sigilo operacional, risco para infra crítica.	Média a alta.	Maior em nuvem remota (dados fora do perímetro/ jurisdição).	Classificação/ minimização de dados, controles de acesso, segregação de ambientes.
Extração de dados de treino (data extraction)	Ataques para extrair trechos memorizados do treino, potencialmente sensíveis.	Exposição de dados sigilosos se o treino incluiu material sensível.	Alta (difícil provar o que foi "memorizado").	Maior em nuvem remota (aumenta incerteza de auditoria).	Evitar dados sensíveis no treino, filtros, monitoramento de consultas.
Inferência de pertinência (membership inference)	Inferir se um registro específico foi usado no treino.	Revela se um incidente, IOC ou documento integrou *datasets* internos.	Alta.	Maior com pouca visibilidade sobre treino/ajuste.	Avaliações de privacidade, restrições de publicação/ acesso.
Extração do modelo (model extraction)	Consultas sistemáticas para reconstruir um modelo aproximado.	Perda de vantagem, replicação de capacidades, engenharia reversa de heurísticas.	Média.	Maior quando o modelo está exposto via API pública sem controles.	*Rate limiting*, detecção de abuso, *watermarking*, políticas de acesso.
Jailbreak / bypass de alinhamento	*Prompts* adversariais para contornar restrições e obter saídas proibidas.	Geração de conteúdo indevido para exploração/ engenharia social, facilitação de TTPs.	Média.	Similar nos dois cenários (piora sem testes/ atualização).	*Red teaming* contínuo, testes adversariais, monitoramento de *drift*.

Continuação ...

Quadro 1 – conclusão

Classe de ataque	Mecanismo de exploração	Impacto potencial	Dificuldade de detecção	Risco relativo	Mitigações
Ferramentas com permissões excessivas (agency/ over-privilege)	LLM aciona ferramentas (scripts, busca) além do necessário; cadeia de ações insegura.	Ações automatizadas indevidas, escalada lateral, impacto em continuidade.	Média a alta.	Maior em nuvem remota (integrações "caixa-preta").	Privilégio mínimo, aprovação humana para ações críticas, *logging* e *rollback*.
Cadeia de suprimentos (supply chain) em software/ IA	Dependências, conectores, bibliotecas, atualizações e artefatos comprometidos.	Comprometimento sistêmico, *backdoors*, indisponibilidade, perda de integridade.	Alta.	Maior quando atualização e dependências são controladas por terceiros.	SSDF, verificação de dependências, gestão de versões, validação pré-produção.

Fonte: Adaptado de OWASP (2024), NIST (2023), Greshake et al. (2023), Biggio; Nelson; Laskov (2012), Steinhardt; Koh; Liang (2017), Shokri et al. (2017), Carlini et al. (2021), Tramèr et al. (2016), Zou et al. (2023), NIST (2006; 2010; 2012; 2022)

O Quadro 2 articula os vetores de dependência a partir de três dimensões: infraestrutura, modelos e governança. As evidências documentais nacionais aqui mobilizadas são, principalmente, a Estratégia Nacional de Cibersegurança (Decreto nº 12.573/2025) e o conjunto PND/END/Livro Branco aprovado pelo Decreto nº 12.725/2025, ambos por sua natureza de diretriz estatal (BRASIL, 2025a; BRASIL, 2025b). Para o risco jurídico-operacional, adiciona-se referência a um dispositivo legal estrangeiro frequentemente citado na literatura de governança de dados em nuvem: o CLOUD Act (UNITED STATES, 2018), por seu papel em explicitar disputas de jurisdição sobre dados armazenados fora do território.

Quadro 2 – Dependência estrutural e tipos de risco (LLMs)

Dimensão	Descrição do vetor	Evidência documental	Tipo de risco associado	Consequência para autonomia estratégica
Infraestrutura	Dependência de computação e serviços de nuvem para inferência/ armazenamento/ integrações	Diretrizes nacionais de cibersegurança e defesa tratam continuidade, resiliência e autonomia como objetivos estruturantes, ainda que não detalhem arquitetura de IA por componente (BRASIL, 2025a; BRASIL, 2025b).	Técnico; Geopolítico	Continuidade sob contestação e capacidade de operar degradado dependem de controle de infraestrutura e planos de contingência (NIST, 2010; BRASIL, 2025a).
Modelos	Dependência de LLMs proprietários (pesos fechados, dados de treino opacos, ajustes e salvaguardas externas)	Estratégias nacionais demandam autonomia tecnológica e redução de vulnerabilidades sistêmicas; isso é tensionado quando o "núcleo cognitivo" é caixa-preta.(BRASIL, 2025b; NIST, 2023).	Técnico; Geopolítico	Auditabilidade limitada, dificuldade de validação independente e menor poder de barganha sobre roadmap, correções e transparência.
Governança	Dependência de termos contratuais, políticas de retenção, logs, telemetria, compliance e jurisdição	Enciber enfatiza governança e segurança; o CLOUD Act explicita que dados sob controle de provedores podem ser objeto de demandas sob jurisdição norte-americana, ainda que armazenados no exterior (BRASIL, 2025a; UNITED STATES, 2018).	Jurídico-operacional; Geopolítico	Risco de incerteza sobre acesso/ retensão e de restrições impostas por terceiros; impacto na integridade e na capacidade de auditoria e responsabilização.

Fonte: Adaptado de Brasil (2025a; 2025b), NIST (2023; 2010) e United States (2018).

O Quadro 3 traduz o argumento em requisitos práticos. Ele não propõe solução ideal, mas um mínimo auditável: controles que permitem reduzir exposição e, sobretudo, tornar falhas rastreáveis, sem o que accountability é apenas uma palavra elegante. A ênfase em logs, resposta a incidentes e continuidade segue guias clássicos do NIST (NIST, 2006; NIST, 2012; NIST, 2010). A seleção de riscos e mitigação para LLMs toma como referência as categorias do OWASP Top 10 for LLM Applications (OWASP, 2024) e as funções de governança do AI RMF (NIST, 2023).

Quadro 3 – Controles mínimos por tipo de risco

Tipo de risco	Controles técnicos/operacionais e de governança (mínimo)	Base documental verificável
Técnico	Logging centralizado e retenção adequada; trilhas de auditoria de prompts/saídas; segregação e privilégio mínimo em ferramentas; validação de entradas e do corpus (RAG); testes adversariais e red teaming contínuos; rollback/versionamento; planos de resposta a incidentes e de continuidade	NIST (2006; 2012; 2010); OWASP (2024); NIST (2023)
Jurídico-operacional	Política de classificação e minimização de dados; requisitos contratuais de retenção/ telemetria/auditoria; delimitação de locais de processamento; cláusulas de notificação de incidente; exigências de evidências para auditoria.	BRASIL (2025a); NIST (2023); UNITED STATES (2018)
Geopolítico	Estratégia de continuidade com alternativas técnicas (redundância e planos de migração); capacidade local de validação e testes; requisitos de transparência e controles de atualização; reduzir lock-in por desenho arquitetural e governança.	BRASIL (2025b); NIST (2010); NIST (2023)

Fonte: Adaptado de NIST (2006; 2010; 2012; 2023), OWASP (2024), Brasil (2025a; 2025b) e United States (2018).

4.1 Interpretação em três movimentos

As evidências organizadas no Quadro 1 permitem uma leitura em três movimentos. O primeiro é que vulnerabilidades técnicas funcionam como expressão concreta de dependência quando a adoção ocorre em regime de caixa-preta. As classes de ataque mais relevantes para ciberdefesa não se reduzem a "engenharia social melhorada"; elas exploram, sobretudo, assimetria de verificação: o operador passa a depender do desempenho e das garantias do provedor, sem dispor de meios equivalentes para auditar insumos, rastrear mudanças de versão, validar integrações e reconstruir a cadeia causal de falhas. Nessa chave, a caixa-preta não é apenas um traço do produto, mas uma condição operacional de risco, compatível com a discussão sobre opacidade e assimetria epistêmica em ecossistemas tecnológicos (PASQUALE, 2015; BURRELL, 2016) e com o enquadramento do NIST de risco em IA como fenômeno sociotécnico dependente de governança ao longo do ciclo de vida (NIST, 2023).

O segundo movimento é que, em aplicações LLM-integradas, o dano pode ocorrer sem controle direto do modelo: basta interferir no que ele lê (por exemplo, bases, páginas, tickets, documentos) ou no modo como o sistema ao redor confia na sua saída e executa ações. É por isso que vetores como injeção indireta de prompt e envenenamento de dados são particularmente relevantes: o alvo efetivo é a interface entre modelo, dados e ferramentas, onde permissões e automação ampliam impacto (Greshake et al., 2023; BIGGIO; NELSON; LASKOV, 2012; STEINHARDT; KOH; LIANG, 2017; OWASP, 2024). Nesses cenários, a ausência de trilhas de auditoria e de telemetria adequada não é detalhe: ela impede atribuição, aprendizado e correção — e, portanto, enfraquece governança operacional (NIST, 2006; NIST, 2012).

O terceiro movimento é que esse quadro técnico se traduz em implicações institucionais e estratégicas. Se disponibilidade, atualização e evidências (logs) dependem de terceiros, parte da continuidade sob crise e da própria responsabilização fica deslocada para fora do controle estatal. Isso torna a governança — entendida

como controle de versão, capacidade de congelamento/reversão, critérios de retenção de dados, acesso a logs, notificação de incidentes e teste contínuo — um requisito de soberania prática, não um anexo administrativo (NIST, 2010; NIST, 2023). No caso brasileiro, as diretrizes de resiliência e proteção de serviços essenciais e infraestruturas críticas autorizam tratar essas exigências como condição de adoção em funções sensíveis (BRASIL, 2025a). A janela de oportunidade, portanto, não é "ter IA", mas ter verificabilidade compatível com risco de infraestrutura crítica: auditabilidade, controle de atualização, logs e telemetria, e red teaming contínuo; quando isso não é exigível por desenho contratual ou arquitetural, a dependência deixa de ser apenas econômica e passa a ser dependência de governança (NIST, 2023; OWASP, 2024).

5 CONCLUSÃO

Este artigo examinou a seguinte questão: a incorporação de LLMs em rotinas de ciberdefesa e, por extensão, na proteção de infraestruturas críticas, amplia vulnerabilidades e compromete a autonomia estratégica de países do Sul Global quando esses sistemas são consumidos como "caixa-preta" e operados em cadeias tecnológicas externas? A resposta, sustentada pelo referencial de dependência, soberania digital e colonialidade (seção 2.2) e pelos artefatos analíticos (seção 4), é afirmativa. A confirmação não depende de postular intenções hostis permanentes de fornecedores; decorre do desenho do ecossistema: opacidade, assimetria de auditoria e governança, e concentração de infraestrutura e serviços digitais em poucos polos, elementos amplamente discutidos na literatura sobre economia digital e fluxos transfronteiriços de dados (UNCTAD, 2021) e no debate sobre decisões algorítmicas como "sociedade da caixa-preta" (PASQUALE, 2015). Em outras palavras: a vulnerabilidade não é apenas um bug; é também a forma social de provisão da capacidade (NIST, 2023).

A hipótese central — de que emergem três ordens de risco associadas à dependência — foi corroborada.

A primeira é técnico-operacional. O Quadro 1 (seção 4.1) organizou vulnerabilidades de aplicações com LLMs (injeção de prompt, vazamento de dados, excesso de permissões em ferramentas, dependências e falhas de integração) e as conectou a cenários plausíveis de uso em SOC/CSIRT e ambientes de infraestrutura crítica, conforme taxonomias abertas voltadas a LLM apps (OWASP, 2024). A leitura desses riscos foi ancorada na concepção sociotécnica de risco em IA e em requisitos de governança, monitoramento e medição propostos pelo AI RMF (NIST, 2023), além de boas práticas consolidadas para resposta a incidentes, gestão de logs e continuidade, necessárias justamente quando um componente crítico passa a mediar triagem, priorização e produção textual da decisão (NIST, 2006; NIST, 2012; NIST, 2010). Nesse quadro, o ganho de produtividade frequentemente associado a LLMs, embora real em tarefas linguísticas, não elimina o requisito de controle: ele desloca o risco para o "entre" — integrações, permissões e ciclos de atualização —, isto é, para o que é menos visível e, por isso, mais perigoso em organizações pressionadas por tempo.

A segunda ordem de risco é jurídico-institucional (accountability). Quando o LLM é consumido como serviço, parte da accountability passa a depender de cláusulas contratuais, transparência mínima de funcionamento, acesso a logs e direitos de auditoria; e esses elementos não são periféricos, pois são as precondições de investigação, correção e aprendizagem institucional após incidentes (NIST, 2012; NIST, 2006). Além disso, a problemática de jurisdição não é abstrata: instrumentos legais estrangeiros podem criar obrigações de preservação e entrega de dados sob determinadas condições, o que precisa ser considerado em políticas de risco e desenho de arquitetura quando dados sensíveis atravessam fronteiras (UNITED STATES, 2018). O ponto analítico é direto: sem verificabilidade independente e sem governança documental e técnica, a responsabilização tende a ficar diluída, e a defesa se torna dependente da boa-fé e da capacidade de prova do próprio provedor, cenário típico de sistemas de caixa-preta (PASQUALE, 2015).

A terceira ordem é geopolítico-estrutural, porque a dependência aqui não é apenas de produto, mas de infraestrutura e de cadeia de valor da economia digital. A literatura de dependência (na sua origem e em releituras) permite interpretar essas assimetrias como padrões persistentes de especialização e controle de capacidades, com efeitos sobre autonomia decisória (PREBISCH, 1950). O aporte da colonialidade, por sua vez, ajuda a nomear um deslocamento específico: não se trata somente de importar tecnologia, mas de operar sob regimes epistêmicos e infraestruturais em que os critérios de verdade, risco aceitável e "forma correta" de governar o sistema são definidos exogenamente, com opacidade e assimetria (QUIJANO, 2000; PASQUALE, 2015). Em um "mundo multiplex", no qual potências médias e emergentes lidam com múltiplos centros de poder e padrões em disputa, a dependência cognitiva/infraestrutural pode atuar, no limite, como vetor de coerção por negação de serviço, restrição de acesso, mudança unilateral de termos ou degradação de desempenho sob contestação — ainda que não seja necessário postular esse desfecho como inevitável para reconhecer a vulnerabilidade estratégica (ACHARYA, 2014; UNCTAD, 2021).

Esses achados tensionam, de forma produtiva, a política brasileira: há diretrizes explícitas de autonomia, soberania e fortalecimento de capacidades nacionais em documentos estruturantes de defesa, que reconhecem a centralidade de tecnologia e indústria na sustentação do poder nacional (BRASIL, 2020a; BRASIL, 2020b; BRASIL, 2025b). Ao mesmo tempo, a aceleração do uso de serviços digitais avançados (incluindo IA) desloca a fronteira do problema: não basta "ter o sistema"; é preciso governar o ciclo de vida, os dados e o regime de atualização (NIST, 2023). Isso abre, no entanto, uma janela de oportunidade: o país pode traduzir diretrizes de autonomia em requisitos objetivos de contratação e arquitetura — exigindo trilhas de auditoria, métricas de monitoramento, controle de versão/atualização, testes contínuos e red teaming, além de governança de desenvolvimento e dependências —, coerentes com boas práticas de engenharia e risco (NIST, 2022; NIST, 2023; OWASP, 2024). Do ponto de vista de governança, a Estratégia Nacional de Cibersegurança fornece um eixo institucional

para tratar resiliência e proteção de ativos críticos em chave nacional, o que pode ser estendido ao uso governado de IA em funções sensíveis, sem confundir modernização com terceirização irrestrita de capacidade (BRASIL, 2025a).

Como agenda de pesquisa, três movimentos são particularmente necessários. Primeiro, a realização de exercícios controlados (sem execução ofensiva real), voltados a medir como vulnerabilidades descritas em taxonomias abertas se manifestam em fluxos operacionais de SOC/CSIRT (OWASP, 2024), acompanhando efeitos em tempo de detecção, qualidade de triagem, integridade de logs e tempos de resposta (NIST, 2006; NIST, 2012). Segundo, a construção de métricas e indicadores de governança para LLMs em ciberdefesa, em linha com a abordagem do AI RMF (NIST, 2023), para tornar auditável aquilo que hoje é tratado como "capacidade intangível". Terceiro, uma comparação sistemática com outros países do Sul Global, para identificar padrões de lock-in, caminhos de mitigação e arranjos híbridos (infraestrutura soberana + componentes externos) que aumentem resiliência sem prometer autossuficiência imediata — comparação que ganha aderência ao dossiê ao tratar autonomia como trajetória institucional, e não como slogan (ACHARYA, 2014; UNCTAD, 2021; PREBISCH, 1950).

REFERÊNCIAS

ACHARYA, Amitav. **The End of American World Order**. Cambridge: Polity Press, 2014.

BIGGIO, Battista; NELSON, Blaine; LASKOV, Klaus. Poisoning attacks against support vector machines. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 29., 2012, Edinburgh. **Proceedings...** New York: ACM, 2012. p. 1467-1474. Disponível em: https://dl.acm.org/doi/10.5555/3042573.3042724. Acesso em: 22 jun. 2026.

BIGGIO, Battista; ROLI, Fabio. Wild patterns: Ten years after the rise of adversarial machine learning. **Pattern Recognition**, Amsterdam, v. 84, p. 317-331, 2018. Disponível em: https://doi.org/10.1016/j.patcog.2018.07.023. Acesso em: 16 jun. 2026.

BRASIL. Decreto nº 12.573, de 4 de agosto de 2025. Institui a Estratégia Nacional de Cibersegurança. **Diário Oficial da União**: seção 1, Brasília, DF, 5 ago. 2025a. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/decreto/D12573.htm. Acesso em: 18 jun. 2026.

BRASIL. Decreto nº 12.725, de 18 de novembro de 2025. Aprova a Política Nacional de Defesa, a Estratégia Nacional de Defesa e o Livro Branco de Defesa Nacional. **Diário Oficial da União**: seção 1, Brasília, DF, 18 nov. 2025b. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/decreto/D12725.htm. Acesso em: 16 jun. 2026.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). **Diário Oficial da União**: seção 1, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Acesso em: 22 jun. 2026.

BRASIL. Ministério da Defesa. **Doutrina Militar de Defesa**: DMiD (MD51-M-04). 3. ed. Brasília, DF: Ministério da Defesa, 2025. Disponível em: https://www.gov.br/defesa/pt-br/assuntos/estado-maior-conjunto-das-forcas-armadas/doutrina-militar/publicacoes-1/publicacoes/md51-m-04-doutrina-militar-de-defesa-dmid-3a-ed-2025.pdf. Acesso em: 18 jun. 2026.

BRASIL. Ministério da Defesa. **Livro Branco de Defesa Nacional**. Brasília, DF: Ministério da Defesa, 2020b. Disponível em: https://www.gov.br/defesa/pt-br/assuntos/copy_of_estado-e-defesa/livro_branco_congresso_nacional.pdf. Acesso em: 05 jun. 2026.

BRASIL. Ministério da Defesa. **Política Nacional de Defesa e Estratégia Nacional de Defesa**. Brasília, DF: Ministério da Defesa, 2020a. Disponível em: https://www.gov.br/defesa/pt-br/assuntos/copy_of_estado-e-defesa/pnd_end_congressonacional_22_07_2020.pdf. Acesso em: 18 jun. 2026.

BURRELL, Jenna. How the machine 'thinks': Understanding opacity in machine learning algorithms. **Big Data & Society**, v. 3, n. 1, p. 1-12, 2016. Disponível em: https://doi.org/10.1177/2053951715622512. Acesso em: 17 jun. 2026.

CARDOSO, Fernando Henrique; FALETTO, Enzo. **Dependência e desenvolvimento na América Latina**: ensaio de interpretação sociológica. 2. ed. Rio de Janeiro: Zahar Editores, 1979.

CARLINI, Nicholas et al. Extracting Training Data from Large Language Models. In: USENIX SECURITY SYMPOSIUM, 30., 2021, Vancouver. **Proceedings...** Berkeley, CA: USENIX Association, 2021. Disponível em: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting. Acesso em: 18 jun. 2026.

COULDRY, Nick; MEJIAS, Ulises A. **The Costs of Connection**: How Data Is Colonizing Human Life and Appropriating It for Capitalism. Stanford: Stanford University Press, 2019.

DOS SANTOS, Theotonio. The Structure of Dependence. **The American Economic Review**, Nashville, v. 60, n. 2, p. 231-236, 1970.

ESTADOS UNIDOS. Department of the Army. U.S. Army Training and Doctrine Command. **TRADOC Pamphlet 525-3-1**: The U.S. Army in Multi-Domain Operations 2028. Fort Eustis, VA: TRADOC, 2018. Disponível em: https://adminpubs.tradoc.army.mil/pamphlets/TP525-3-1.pdf. Acesso em: 02 jun. 2026.

ESTADOS UNIDOS. Department of Defense. **Summary of the Joint All-Domain Command and Control (JADC2) Strategy**. Washington, DC: DoD, 2022. Disponível em: https://media.defense.gov/2022/Mar/17/2002958406/-1/-1/1/SUMMARY-OF-THE-JOINT-ALL-DOMAIN-COMMAND-AND-CONTROL-STRATEGY.PDF. Acesso em: 18 jun. 2026.

FARRELL, Joseph; KLEMPERER, Paul. Coordination and Lock-In. *In*: ARMSTRONG, Mark; PORTER, Robert H. (ed.). **Handbook of Industrial Organization**. Amsterdam: Elsevier, 2007. v. 3, cap. 31, p. 1973-2025.

GEBRU, Timnit et al. Datasheets for Datasets. **arXiv preprint arXiv:1803.09010**, 2018. Disponível em: https://arxiv.org/abs/1803.09010. Acesso em: 02 jun. 2026.

GRESHAKE, Kai et al. Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. **arXiv**, 2023. Disponível em: https://arxiv.org/abs/2302.12173. Acesso em: 05 jun. 2026.

ISSMAEL JÚNIOR, Ali Kamel. Operações Multidomínio (MDO) no Sul Global: modelo mínimo viável para C2 e dados. **InterAção**, Santa Maria, v. 17, n. 2, e95275, p. 1-24, jun. 2026. DOI: https://doi.org/10.5902/2357797595275. Acesso em: 10 jun. 2026.

MITCHELL, Margaret et al. Model Cards for Model Reporting. *In*: CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2019, Atlanta. **Proceedings...** New York: ACM, 2019. p. 220-229. Disponível em: https://dl.acm.org/doi/10.1145/3287560.3287596. Acesso em: 05 jun. 2026.

NEIVA FILHO, Ivan Ferreira; SILVA, Karen Cristina Leal da. Base Industrial de Defesa Brasileira: vetor estruturante da soberania, da inovação tecnológica e do aumento da competitividade industrial. **InterAção**, Santa Maria, v. 17, n. 2, e96176, p. 1-30, jun. 2026. DOI: https://doi.org/10.5902/2357797596176. Acesso em: 09 jun. 2026.

NIST. **Artificial Intelligence Risk Management Framework (AI RMF 1.0)**. Gaithersburg, MD: National Institute of Standards and Technology, 2023. Disponível em: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf. Acesso em: 08 jun. 2026.

NIST. **Computer Security Incident Handling Guide (SP 800-61 Rev. 2)**. Gaithersburg, MD: National Institute of Standards and Technology, 2012. Disponível em: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-61r2.pdf. Acesso em: 18 jun. 2026.

NIST. **Contingency Planning Guide for Federal Information Systems (SP 800-34 Rev. 1)**. Gaithersburg, MD: National Institute of Standards and Technology, 2010. Disponível em: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-34r1.pdf. Acesso em: 11 jun. 2026.

NIST. **Guide to Computer Security Log Management (SP 800-92)**. Gaithersburg, MD: National Institute of Standards and Technology, 2006. Disponível em: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-92.pdf. Acesso em: 18 jun. 2026.

NIST. **Guide to Industrial Control Systems (ICS) Security** (SP 800-82 Rev. 2). Gaithersburg, MD: National Institute of Standards and Technology, 2015. Disponível em: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-82r2.pdf. Acesso em: 22 jun. 2026.

NIST. **Managing Information Security Risk**: Organization, Mission, and Information System View (SP 800-39). Gaithersburg, MD: National Institute of Standards and Technology, 2013. Disponível em: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-39.pdf. Acesso em: 22 jun. 2026.

NIST. **Secure Software Development Framework (SSDF) Version 1.1 (SP 800-218)**. Gaithersburg, MD: National Institute of Standards and Technology, 2022. Disponível em: https://csrc.nist.gov/pubs/sp/800/218/final. Acesso em: 10 jun. 2026.

ORGANIZAÇÃO DO TRATADO DO ATLÂNTICO NORTE (OTAN). **AJP-3.10**: Allied Joint Doctrine for Electronic Warfare. Bruxelas: OTAN, 2015. Disponível em: https://mpsotc.army.gr/wp-content/uploads/2024/03/2.-AJP-3.10-EDA-V1-E.pdf. Acesso em: 18 jun. 2026.

ORGANIZAÇÃO DO TRATADO DO ATLÂNTICO NORTE (OTAN). **AJP-3.20**: Allied Joint Doctrine for Cyberspace Operations. Bruxelas: OTAN, 2018. Disponível em: https://iwar.org.uk/wp-content/uploads/2021/06/AJP-3.20-EDA-V1-E.pdf. Acesso em: 10 jun. 2026.

OWASP. **OWASP Top 10 for Large Language Model Applications**. OWASP Foundation, 2024. Disponível em: https://owasp.org/www-project-top-10-for-large-language-model-applications/. Acesso em: 02 jun. 2026.

PASQUALE, Frank. **The Black Box Society**: The Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press, 2015.

POCHMANN, Pablo Gustavo Cogo. A revolução da simulação construtiva no Sul Global: comparativo do uso do software SWORD. **InterAção**, Santa Maria, v. 17, n. 2, e96435, p. 1-15, jun. 2026. DOI: https://doi.org/10.5902/2357797596435. Acesso em: 18 jun. 2026.

PREBISCH, Raúl. **The Economic Development of Latin America and Its Principal Problems**. New York: United Nations, 1950. Disponível em: https://repositorio.cepal.org/handle/11362/29973. Acesso em: 18 jun. 2026.

QUIJANO, Aníbal. Coloniality of Power, Eurocentrism, and Latin America. **Nepantla: Views from South**, Durham, v. 1, n. 3, p. 533-580, 2000. Disponível em: https://doi.org/10.1215/15314201-1-3-533. Acesso em: 09 jun. 2026.

SHOKRI, Reza et al. Membership Inference Attacks Against Machine Learning Models. *In*: IEEE SYMPOSIUM ON SECURITY AND PRIVACY, 38., 2017, San Jose. **Proceedings...** Piscataway, NJ: IEEE, 2017. p. 3-18. Disponível em: https://ieeexplore.ieee.org/document/7958568. Acesso em: 22 jun. 2026.

STEINHARDT, Jacob; KOH, Pang Wei; LIANG, Percy. Certified defenses for data poisoning attacks. *In*: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS), 2017. **Proceedings...** 2017. Disponível em: https://arxiv.org/abs/1706.03691. Acesso em: 18 jun. 2026.

TOUVRON, Hugo et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. **arXiv preprint arXiv:2307.09288**, 2023. Disponível em: https://arxiv.org/abs/2307.09288. Acesso em: 22 jun. 2026.

TRAMÈR, Florian et al. Stealing Machine Learning Models via Prediction APIs. *In*: USENIX SECURITY SYMPOSIUM, 25., 2016, Austin. **Proceedings...** Berkeley, CA: USENIX Association, 2016. p. 601-614. Disponível em: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer. Acesso em: 22 jun. 2026.

UNCTAD. **Digital Economy Report 2021**: Cross-border Data Flows and Development: For whom the data flow. Geneva: United Nations Conference on Trade and Development, 2021. Disponível em: https://unctad.org/publication/digital-economy-report-2021. Acesso em: 18 jun. 2026.

UNITED STATES. Clarifying Lawful Overseas Use of Data Act (CLOUD Act). Washington, DC: U.S. Congress, 2018. Disponível em: https://www.congress.gov/bill/115th-congress/house-bill/4943. Acesso em: 20 jun. 2026.

ZOU, Andy et al. Universal and Transferable Adversarial Attacks on Aligned Language Models. **arXiv preprint arXiv:2307.15043**, 2023. Disponível em: https://arxiv.org/abs/2307.15043. Acesso em: 22 jun. 2026.

Autoria

1 Onildo Ribeiro de Assis II

Analista de Sistemas; Administrador; Professor

https://orcid.org/0000-0002-8884-1378 • onildo.ribeiro@gmail.com

2 Ana Isabella Bezerra Lau Ribeiro

Advogada; Mestre pela Universidade Federal da Paraíba; Professora

https://orcid.org/0000-0002-5441-4183 • anaisabellalau@gmail.com

Como citar este artigo

ASSIS II, O. R.; RIBEIRO, A. I. B. L. Ciberdefesa e LLMs no Sul Global: governança, vulnerabilidades e autonomia. **InterAção**, Santa Maria, v. 17, n. 3, e97309, p. 1-34, 2026. DOI 10.5902/2357797597309. Disponível em: https://dx.doi.org/10.5902/2357797597309. Acesso em: dia mês abreviado. ano.