ISSN on-line: 2237-2113
https://estacio.periodicoscientificos.com.br/index.php/destarte
DESTARTE, v.15, n.1, p. 68-90, jan./jun. 2026.
DOI: 10.5281/zenodo.21285122

ARTIGO ORIGINAL

GOVERNANÇA DE TI EM MODELOS GENERATIVOS DE TERCEIROS EXECUTADOS LOCALMENTE: ATUALIZAÇÃO, TELEMETRIA E ROLLBACK EM AMBIENTES CORPORATIVOS DE ALTO RISCO CIBERNÉTICO

ORIGINAL ARTICLE

IT GOVERNANCE FOR LOCALLY EXECUTED THIRD-PARTY GENERATIVE MODELS: UPDATING, TELEMETRY AND ROLLBACK IN HIGH-RISK CORPORATE CYBER ENVIRONMENTS

Ângelo Machado de Souza[1]
Juliano Araújo Santana[2]
Davis Souza Alves[3]
Florida Christian University – FCU, Orlando-USA

Márcio Magera Conceição[4]
Universidade Guarulhos – UNG, Brasil

Resumo

Este artigo desenvolve uma revisão bibliográfica e documental direcionada sobre a governança de inteligência artificial aplicada à gestão da informação e da tecnologia. O foco recai sobre modelos generativos de terceiros executados localmente em ambientes corporativos de alto risco cibernético. O objetivo central é propor um framework conceitual capaz de articular atualização, telemetria e rollback como critérios para admissão, operação e retirada segura de artefatos de inteligência artificial, convertendo a noção ampla de ciclo de vida em cadeia de evidências auditáveis. A pesquisa adota abordagem qualitativa, exploratória e de natureza teórico-conceitual, baseada na análise de 24 documentos científicos, normativos, regulatórios e técnicos primários. Esses materiais foram organizados em seis eixos analíticos: governança de IA; cadeia de suprimento e proveniência; atualização; telemetria e observabilidade; rollback e incidentes; e regulação e jurisdição. Como resultado, consolida-se uma arquitetura estruturada em quatro camadas, acompanhada de uma matriz decisória de riscos e controles mínimos voltada ao uso em gestão, informação e tecnologia da informação. A contribuição mais relevante está em transformar recomendações dispersas em critério operacional auditável e teoricamente situado, com especial utilidade para organizações que precisam controlar a proveniência dos artefatos, os fluxos externos associados e a capacidade efetiva de reversão.

Palavras-chave: governança de IA; gestão da informação; tecnologia da informação; modelos generativos; telemetria e rollback.

---

[1] Mestrando em Administração pela Florida Christian University com pesquisa em Novas Tecnologias Avançadas. MBA em Gerenciamento de Projetos de TI. E-mail: angelomsouza2@gmail.com.
[2] Mestrando em Administração com pesquisa em Inteligência Artificial Explicável (XAI), pela Florida Christian University, pós-graduado em Gestão de TI e Cloud Computing (UFSCar).
[3] Doutor em Administração de TI - Ph.D pela Florida Christian University (EUA) convalidado no Brasil, Mestre em Administração com foco em TI Verde (2015); Atua nos Estados Unidos como Gerente de Projetos de Cibersegurança (R&D) com foco em Privacidade de Dados (LGPD/GDPR), Computação Forense, Ethical Hacker e Inteligência Artificial(AI). E-mail: davis.alves@anppd.org.
[4] Doutor em Economista pela PUC - Campinas. MBA de Marketing – ESAMC, Sorocaba. Mestrado em Administração pela UNG -Guarulhos. Pró-Reitor da Universidade Guarulhos.

Submetido em 25/04/2026
Aceito em 14/05/2026

Abstract

This paper presents a targeted bibliographic and documentary review on AI governance applied to information and technology management, with attention to locally executed third-party generative models in corporate environments exposed to high cyber risk. It proposes a conceptual framework that connects updating, telemetry and rollback as criteria for admission, operation and safe retirement of artificial intelligence artifacts, translating the broad notion of lifecycle governance into a chain of auditable evidence. The method is qualitative, exploratory and theoretical-conceptual, based on the analysis of 24 scientific, normative, regulatory and primary technical documents organized into six axes: AI governance, supply chain and provenance, updating, telemetry and observability, rollback and incidents, and regulation and jurisdiction. The study consolidates a four-layer architecture and a decision matrix of risks and minimum controls for management, information and information technology decisions. Its main contribution is to convert dispersed recommendations into a theoretically situated and auditable operational criterion for environments that need to control provenance, external flows and actual rollback capability.

Keywords: AI governance; information management; information technology; generative models; telemetry and rollback.

INTRODUÇÃO

A adoção de modelos generativos em organizações tem estimulado arquiteturas locais ou controladas pela própria empresa, frequentemente justificadas por latência, custo, confidencialidade, exigências contratuais e dependência reduzida de provedores remotos. Entretanto, a execução local não elimina riscos de governança; antes, desloca a decisão para a proveniência do artefato, as dependências técnicas, os canais de atualização, os fluxos de telemetria, a capacidade de monitoramento e a reversão segura. Nessa perspectiva, a governança de IA deve ser compreendida como processo de ciclo de vida, e não como avaliação isolada no momento da aquisição ou instalação do modelo. Para que essa compreensão não permaneça apenas como princípio geral de gestão, ela precisa ser traduzida em evidências operacionais verificáveis (Tabassi, 2023; ISO/IEC, 2023a; ISO/IEC, 2023b).

Em ambientes corporativos sensíveis, a inferência local pode introduzir problemas específicos: mudanças silenciosas de versão, formatos de serialização inseguros, dependências não congeladas, suporte remoto, sincronização automática, persistência de logs, fluxos externos acessórios e exposição regulatória associada a dados ou metadados. Esses fatores são particularmente relevantes quando modelos de terceiros passam a apoiar análise de documentos, geração de relatórios, triagem

informacional, apoio operacional ou automação em setores sujeitos a risco cibernético elevado.

A literatura e os guias técnicos sobre IA generativa, desenvolvimento seguro, cadeia de suprimento, implantação segura e monitoramento pós-implantação indicam que a segurança de um sistema de IA não depende apenas da qualidade preditiva do modelo. Governar esse tipo de artefato implica conhecer o que foi incorporado, sua origem, o modo como é atualizado, os componentes que o acompanham, os fluxos externos permitidos e as condições em que a organização consegue reverter uma versão comprometida ou incompatível. Ao articular essas dimensões, o estudo assume que sua contribuição teórica não está em criar uma nova classe de controles, mas em tensionar frameworks gerais de governança diante de uma situação operacional específica: a execução local de modelos generativos de terceiros em ambientes sensíveis (Autio et al., 2024; Batool et al., 2025).

A questão de pesquisa é: quais arranjos organizacionais de atualização, telemetria e rollback podem reduzir vazamento informacional, desvio operacional e risco regulatório na execução local de modelos generativos de terceiros? O objetivo geral é propor um framework conceitual de governança técnica para orientar admissão, implantação, operação e retirada segura desses modelos. Como objetivos específicos, o estudo busca delimitar o objeto de confiança, identificar vetores de risco recorrentes, organizar evidências em torno da tríade atualização-telemetria-rollback e converter essa síntese em uma matriz decisória aplicável a engenharia, computação e tecnologia da informação.

O escopo do trabalho concentra-se em modelos generativos de terceiros executados em infraestrutura local, ambiente on premises, nuvem privada ou ambiente controlado pela organização. Não são objeto central da análise o treinamento próprio de modelos fundacionais, o consumo exclusivamente em software como serviço, a avaliação comparativa de desempenho de modelos ou o debate ético desvinculado de controles técnicos e de gestão de risco. A contribuição está na integração de recomendações que aparecem dispersas em documentos científicos, técnicos, normativos e regulatórios, oferecendo uma leitura operacional para decisões de governança em ambientes de alto risco.

FUNDAMENTAÇÃO TEÓRICA E CONTEXTUALIZAÇÃO

GOVERNANÇA DE IA COMO CICLO DE VIDA

A governança de IA é tratada, em normas e frameworks recentes, como um processo contínuo de identificação, avaliação, tratamento e monitoramento de riscos. O AI RMF 1.0 estrutura funções de governança, mapeamento, mensuração e gestão, enquanto o perfil de IA generativa reforça a necessidade de monitoramento, gestão de incidentes, documentação e avaliação contextual de riscos (Tabassi, 2023; Autio et al., 2024). A convergência entre esses documentos desloca a governança de uma decisão pontual para uma prática de ciclo de vida. Ainda assim, funções e categorias amplas nem sempre especificam quais evidências técnicas tornam defensável a admissão, a atualização ou a retirada de um modelo generativo de terceiro executado localmente. Essa leitura é compatível com normas internacionais de gestão de riscos e sistemas de gestão de IA, nas quais responsabilidades, processos, evidências e revisão contínua são elementos necessários para que a organização não trate a IA como ferramenta isolada (ISO/IEC, 2023a; ISO/IEC, 2023b).

A literatura de governança também destaca que a IA deve ser compreendida como sistema sociotécnico: dados, modelo, infraestrutura, usuários, processos, contratos e políticas institucionais integram a superfície decisória (Batool et al., 2025). Essa contribuição amplia a análise para além do artefato técnico, mas também apresenta uma tensão relevante para este estudo: quanto mais abrangente se torna a noção de governança, maior o risco de diluição dos controles concretos necessários à operação local. No caso de modelos generativos de terceiros executados localmente, essa interpretação é especialmente importante porque a organização passa a controlar partes da operação que, no consumo remoto, podem permanecer sob responsabilidade primária do provedor. A mudança de arquitetura, portanto, não reduz necessariamente o risco; ela redistribui responsabilidades e exige evidências mais próximas da engenharia operacional.

O tensionamento entre esses autores permite delimitar a posição adotada neste estudo. Tabassi (2023) e Autio et al. (2024) oferecem uma gramática de ciclo de vida e gestão de riscos; ISO/IEC 23894:2023 e ISO/IEC 42001:2023 reforçam processos, responsabilidades e gestão contínua; Batool, Zowghi e Bano (2025) ampliam a governança para o plano sociotécnico. O ponto ainda em aberto está na passagem desses enquadramentos para a operação local de modelos generativos de terceiros, isto é, em como converter princípios de governança em critérios auditáveis sobre atualização, telemetria e rollback.

A contribuição teórica deste estudo, portanto, é tratar o ciclo de vida não apenas como sequência de fases, mas como cadeia de evidências. Cada decisão - admitir, implantar, atualizar, monitorar, reverter ou retirar - deve estar vinculada a registros mínimos sobre origem, integridade, dependências, fluxos informacionais, responsáveis e reversibilidade. A lacuna enfrentada situa-se na passagem entre frameworks gerais de governança de IA e requisitos técnico-operacionais verificáveis para ambientes corporativos de alto risco cibernético. Essa formulação evita duas reduções: a de uma governança apenas normativa, sem lastro operacional, e a de uma segurança apenas técnica, sem responsabilização organizacional.

CADEIA DE SUPRIMENTO, DESENVOLVIMENTO SEGURO E IMPLANTAÇÃO

A cadeia de suprimento digital amplia a noção de objeto de confiança. Artefatos tecnológicos devem ser avaliados pela origem, integridade, dependências, mecanismos de atualização e capacidade de resposta a falhas ou vulnerabilidades (Boyens et al., 2024). Quando essa lógica é aplicada à IA generativa, o objeto de confiança deixa de ser apenas o arquivo de pesos e passa a incluir tokenizadores, formatos de serialização, utilitários de conversão, imagens de contêiner, bibliotecas de carregamento, caches, conectores, políticas de rede e documentação associada.

No eixo de desenvolvimento seguro, o SSDF e o perfil comunitário para IA generativa e modelos fundacionais reforçam práticas de mitigação de vulnerabilidades, rastreabilidade, gestão de componentes e segurança ao longo do ciclo de desenvolvimento e implantação (Souppaya; Scarfone; Dodson, 2022; Booth et al., 2024). Tais práticas são relevantes para o artigo porque atualizações de

modelos e bibliotecas não devem ser tratadas como simples substituição de arquivos. A promoção de versão é uma mudança técnica com impacto sobre integridade, comportamento, compatibilidade, superfície de ataque e obrigações de registro.

Guias institucionais sobre segurança de sistemas de IA indicam a necessidade de due diligence, tratamento seguro de pesos serializados, registro, monitoramento, resposta a incidentes e rollback como elementos de operação resiliente (NCSC et al., 2023; AISC et al., 2024). O guia de implantação segura é particularmente aderente ao problema aqui analisado porque considera organizações que implantam e operam sistemas de IA desenvolvidos externamente em ambientes on premises ou em nuvem privada, especialmente em contextos de alto valor e alta ameaça (AISC et al., 2024).

Além disso, recomendações específicas sobre riscos de cadeia de suprimento de IA e aprendizado de máquina reforçam que a contratação ou incorporação de componentes de terceiros deve orientar perguntas e requisitos para fornecedores, cobrindo riscos e mitigações em desenvolvimento, aquisição e operação (ACSC, 2024; ACSC, 2026). A literatura recente sobre cadeia de suprimento de LLMs confirma que riscos de proveniência, reutilização de componentes, dependências e integrações permanecem problemas abertos de segurança em ecossistemas que combinam modelos, datasets, bibliotecas e ferramentas de terceiros (Hu et al., 2024; Wang et al., 2024).

TELEMETRIA, OBSERVABILIDADE E EXPOSIÇÃO INFORMACIONAL

Nesse contexto, a telemetria requer distinção entre observabilidade interna e saída externa. Logs, métricas, trilhas de auditoria e alertas são necessários para a operação segura, mas não se confundem com envio de dados ou metadados a fornecedores por mecanismos de analytics, verificação automática de versões, autenticação, sincronização ou suporte. Em ambientes sensíveis, a pergunta crítica não é apenas se o modelo é local, mas se a pilha de execução realiza chamadas

Destarte, v.15, n.1, p. 68-90, jan./jun. 2026 DOI: 10.5281/zenodo.21285122

externas, grava logs sensíveis ou reativa dependências que alteram o perímetro informacional.

Documentações técnicas de ecossistemas utilizados em IA mostram que variáveis de ambiente, modo offline, uso de arquivos locais e semântica de serialização podem interferir diretamente na segurança operacional. A existência de recursos de controle de telemetria, modo offline ou carregamento local não elimina o risco por si só; ela exige configuração, validação e evidência verificável no ambiente em que o artefato será executado (HUGGING FACE, s.d.a; HUGGING FACE, s.d.b; PYTORCH, s.d.).

A literatura sobre segurança e privacidade em LLMs e taxonomias aplicadas de risco reforça ameaças como vazamento de prompts, logs, dados sensíveis, segredos operacionais, extração de modelos e extração de dados de treinamento (Yao et al., 2024; OWASP Foundation, 2025). Estudos clássicos e recentes indicam que modelos podem estar sujeitos a extração por interfaces de predição e a divulgação indevida de dados memorizados, o que justifica controles sobre retenção, acesso, segregação e uso de artefatos (Tramèr et al., 2016; Carlini et al., 2023).

A dimensão regulatória também afeta arquiteturas locais. Fluxos acessórios, suporte remoto, telemetria de fornecedor ou sincronização podem reativar obrigações relacionadas à transferência internacional de dados, papéis dos agentes e bases normativas. No contexto brasileiro, a Resolução CD/ANPD n. 19/2024 regula transferência internacional de dados; no âmbito europeu, o Regulamento de IA estabelece obrigações por papéis e riscos para sistemas de IA (Autoridade Nacional de Proteção de Dados, 2024; União Europeia, 2024). Por isso, a governança técnica deve dialogar com privacidade, jurídico e gestão de fornecedores.

ARQUITETURA E TAXONOMIA DO PROBLEMA

Em continuidade à revisão bibliográfica e documental, esta seção organiza a arquitetura conceitual do problema antes da proposição do framework. A taxonomia adotada deriva da leitura do corpus e delimita três dimensões operacionais - atualização, telemetria e rollback -, articuladas a um objeto de confiança ampliado. O

Destarte, v.15, n.1, p. 68-90, jan./jun. 2026 DOI: 10.5281/zenodo.21285122

propósito é explicitar quais elementos compõem a decisão técnica quando uma organização avalia um modelo generativo de terceiro para execução local.

O objeto de confiança é entendido como a composição mínima de artefatos, dependências e fluxos necessários para que o modelo opere no ambiente corporativo. Essa composição inclui o arquivo do modelo, pesos, tokenizadores, configuração, bibliotecas, imagens, scripts de conversão, runtime, logs, integrações, políticas de rede, variáveis de ambiente, documentação de licença e registros de mudança. O quadro 1 sintetiza esse objeto de confiança, cuja verificação sustenta a rastreabilidade ao longo da vida do sistema.

Quadro 1 - Taxonomia do objeto de confiança em modelos generativos de terceiros executados localmente

Elemento	Descrição operacional	Risco associado
Artefato principal	Pesos, checkpoint, configuração e tokenizador efetivamente incorporados ao ambiente.	Artefato malicioso, incompatível, adulterado ou sem proveniência verificável.
Dependências de execução	Bibliotecas, runtime, imagens, utilitários de conversão e pacotes auxiliares.	Reativação de vulnerabilidades, incompatibilidade e mudança de comportamento após atualização.
Fluxos informacionais	Chamadas externas, sincronização, analytics, suporte, autenticação, logs e metadados.	Saída não mapeada de dados, prompts, segredos operacionais ou metadados.
Controles operacionais	Política de rede, logs internos, baseline, revisão periódica, segregação e aprovação de mudança.	Ausência de evidência para auditoria e resposta a incidentes.
Reversibilidade	Versões anteriores preservadas, gatilhos formais, playbook e ensaio de recuperação.	Persistência de versão comprometida ou impossibilidade prática de rollback.
Conformidade	Mapeamento de papéis, bases normativas, transferência de dados e obrigações contratuais.	Exposição jurisdicional e obrigações regulatórias não avaliadas.

Fonte: elaborado pelos autores (2026), com base na síntese do corpus documental.

A atualização é interpretada como mudança governada. Uma nova versão pode corrigir falhas, mas também alterar comportamento, dependências, requisitos de hardware, chamadas externas, políticas de cache ou compatibilidade com controles internos. A telemetria, por sua vez, é tratada como disciplina de fluxo: observa-se o sistema internamente, enquanto se mapeia e limita qualquer comunicação externa. O rollback corresponde à capacidade material de desfazer

uma decisão de implantação, com integridade, gatilhos e responsabilidades definidos.

A taxonomia permite formular uma regra de decisão: um modelo generativo de terceiro executado localmente não deve ser avaliado apenas por funcionalidade, acurácia percebida ou conveniência operacional. Em ambientes sensíveis, a admissão depende de evidências sobre proveniência, licença, integridade, dependências, fluxos externos, registros, responsáveis e reversibilidade. Essa regra orienta a revisão do estado da arte e sustenta a estrutura proposta nas seções seguintes.

PROCEDIMENTOS METODOLÓGICOS DA REVISÃO DOCUMENTAL

O estudo é qualitativo, exploratório e teórico-conceitual, complementado por revisão bibliográfica e documental direcionada. A revisão tem função analítica, não bibliométrica, e não se apresenta como revisão sistemática ou meta-análise. Essa opção metodológica é adequada ao objetivo do trabalho porque o problema investigado não exige mensuração de efeito, comparação estatística entre estudos ou exaustão bibliométrica da literatura, mas a organização crítica de evidências heterogêneas para sustentar um framework conceitual de governança técnica. A decisão decorre da natureza do objeto: atualização, telemetria, operação offline, serialização, cadeia de suprimento e rollback aparecem distribuídos em normas, guias institucionais, documentação técnica primária, regulação e literatura científica. Por isso, o valor do procedimento está menos na representatividade estatística do corpus e mais na capacidade de articular documentos de naturezas distintas em torno de problemas operacionais recorrentes.

A busca foi realizada entre 18 e 21 de março de 2026 em três grupos de fontes: repositórios normativos e institucionais; documentação técnica primária; e literatura científica e documentos de pesquisa. O rastreamento inicial recuperou 34 itens candidatos. Em seguida, a leitura de escopo removeu duplicidades e documentos sem aderência suficiente ao recorte. O corpus final foi composto por 24 documentos, selecionados por sua relação direta com ciclo de vida de IA, segurança, privacidade, operação local, atualização, observabilidade, incidentes, regulação ou jurisdição. Foram incluídos documentos científicos, normativos, regulatórios ou

técnicos primários que contribuíam para pelo menos um dos eixos analíticos; foram excluídos textos promocionais, notícias sem característica documental, opiniões sem instrumental técnico mínimo e estudos voltados apenas ao desempenho de modelos. A seleção privilegiou documentos capazes de oferecer contribuição analítica para a formulação do framework, e não a exaustão bibliométrica do campo.

Os documentos foram classificados em seis eixos: governança de IA; cadeia de suprimento e proveniência; atualização e mudança; telemetria e observabilidade; rollback e incidentes; e regulação e jurisdição. O critério de suficiência do corpus foi definido pela reiteração funcional das categorias: novas buscas passaram a repetir funções analíticas já cobertas - admissão, implantação, operação, atualização, disciplina de fluxos, reversibilidade, incidentes e conformidade - sem acrescentar categoria relevante para a decisão sobre atualização, telemetria ou rollback. Esse critério não equivale à saturação estatística nem à exaustão da literatura; trata-se de suficiência argumentativa para uma síntese conceitual. Os critérios de seleção foram, portanto, funcionais: cada documento permaneceu no corpus quando contribuía para pelo menos um eixo analítico e para a formulação de critérios de governança aplicáveis à tríade atualização-telemetria-rollback. A análise seguiu quatro movimentos: marcação de evidências por eixo; extração de proposições úteis para tomada de decisão corporativa; reorganização dos achados pela tríade; e formulação da arquitetura conceitual em camadas. Os quadros 3 e 4 sintetizam o protocolo e a matriz de extração, sustentados por mais de uma família documental.

Quadro 2 - Corpus documental nominal da revisão direcionada

Documento ou fonte	Tipo de documento	Eixo analítico principal	Função no artigo
Tabassi (2023)	Framework normativo	Governança de IA	Define a gramática de ciclo de vida e gestão de riscos.
Autio et al. (2024)	Perfil institucional	Governança de IA generativa	Adapta o AI RMF à avaliação contextual de riscos generativos.
ISO/IEC 23894:2023	Norma internacional	Gestão de riscos de IA	Reforça processos de identificação, avaliação e tratamento contínuo de riscos.
ISO/IEC 42001:2023	Norma internacional	Sistema de gestão de IA	Apoia a definição de responsabilidades, evidências e revisão contínua.
Batool, Zowghi e Bano (2025)	Literatura científica	Governança sociotécnica	Amplia a análise para dados, infraestrutura, usuários, contratos e políticas.
Boyens et al. (2024)	Framework institucional	Cadeia de suprimento e proveniência	Orienta a avaliação de origem, integridade, dependências e riscos de fornecedores.
Souppaya, Scarfone e Dodson (2022)	Framework técnico	Desenvolvimento seguro e atualização	Sustenta práticas de rastreabilidade, mitigação de vulnerabilidades e gestão de componentes.

Documento ou fonte	Tipo de documento	Eixo analítico principal	Função no artigo
Booth et al. (2024)	Perfil técnico	IA generativa e desenvolvimento seguro	Relaciona SSDF a modelos fundacionais e artefatos de IA generativa.
NCSC et al. (2023)	Guia institucional	Implantação segura e incidentes	Reforça due diligence, monitoramento, registro e resposta a incidentes.
AISC et al. (2024)	Guia institucional	Implantação, rollback e resiliência	Apoia controles de operação segura, recuperação e rollback.
ACSC (2024)	Guia institucional	Engajamento com IA e fornecedores	Orienta riscos e perguntas de governança na adoção de IA.
ACSC (2026)	Guia institucional	Cadeia de suprimento de IA/ML	Organiza riscos e mitigações para componentes de terceiros.
Hu et al. (2024)	Artigo de pesquisa	Cadeia de suprimento de LLMs	Identifica problemas abertos de segurança em ecossistemas de LLMs.
Wang et al. (2024)	Artigo de pesquisa	Cadeia de suprimento de LLMs	Apoia a agenda de riscos em modelos, datasets, bibliotecas e ferramentas.
Hugging Face - Environment variables	Documentação técnica	Telemetria e operação offline	Informa controles de variáveis de ambiente, telemetria e modo offline.
Hugging Face - Installation	Documentação técnica	Execução local e atualização	Apoia a discussão sobre instalação, cache, arquivos locais e operação offline.
PyTorch - Serialization semantics	Documentação técnica	Serialização e integridade	Sustenta a atenção a formatos de carregamento, pesos e artefatos serializados.
OWASP Foundation (2025)	Taxonomia de risco	Segurança e privacidade em LLMs	Mapeia ameaças aplicáveis a vazamento, logs, prompts, segredos e extração.
Yao et al. (2024)	Artigo de pesquisa	Segurança e privacidade em LLMs	Sintetiza ameaças de privacidade, segurança e exposição de dados em LLMs.
Tramèr et al. (2016)	Artigo de pesquisa	Extração de modelos	Justifica controles sobre interfaces, acesso e proteção contra extração.
Carlini et al. (2023)	Artigo de pesquisa	Memorização e extração de dados	Reforça riscos de divulgação indevida de dados a partir de modelos.
ANPD (2024)	Regulação	Transferência internacional de dados	Enquadra fluxos acessórios, suporte remoto e transferência de dados no contexto brasileiro.
União Europeia (2024)	Regulação	Regulação e jurisdição de IA	Apoia o mapeamento de papéis, obrigações e riscos regulatórios.
Rao et al. (2026)	Documento técnico	Monitoramento pós-implantação	Evidencia desafios de terminologia, método e acompanhamento de IA em produção.

Fonte: elaborado pelos autores (2026), com base na composição final do corpus documental.

Quadro 3 - Protocolo sintético de busca, seleção e análise

Etapa	Descrição	Função no artigo
Fontes consultadas	NIST, ISO/IEC, ACSC, NCSC, AISC, ANPD, EUR-Lex, Hugging Face, PyTorch, OWASP, periódicos científicos e arXiv.	Ampliar a cobertura entre normas, guias, regulação, documentação técnica e literatura científica.
Descritores nucleares	AI risk management framework; generative AI profile; secure AI deployment; post-deployment monitoring; model supply chain; telemetry; offline mode; serialization semantics; LLM security; model extraction.	Recuperar documentos diretamente associados a governança, cadeia de suprimento, telemetria, atualização e reversão.
Critérios de inclusão	Documentos científicos, normativos, regulatórios ou técnicos primários relacionados ao ciclo de vida de IA, segurança, privacidade, operação local, atualização, observabilidade ou	Manter documentos com função analítica clara para construção do framework.

Etapa	Descrição	Função no artigo
	incidentes.	
Critérios de exclusão	Textos promocionais, notícias sem característica documental, opiniões sem instrumental técnico mínimo e estudos focados apenas em desempenho de modelo.	Evitar fontes sem contribuição metodológica, técnica ou normativa para o problema.
Eixos analíticos	Governança de IA; cadeia de suprimento e proveniência; atualização; telemetria e observabilidade; rollback e incidentes; regulação e jurisdição.	Classificar evidências e extrair proposições úteis para decisão corporativa.
Transparência do escopo	O delineamento explicita período de busca, grupos de fontes, rastreamento inicial, leitura de escopo, critérios funcionais de seleção e composição final do corpus de 24 documentos.	Tornar claros o alcance e os limites da revisão direcionada, sem convertê-la em revisão sistemática ou meta-análise.

Fonte: elaborado pelos autores (2026), com base no protocolo de revisão documental.

Quadro 4 - Matriz analítica de extração por eixo

Eixo analítico	Pergunta de extração	Resultado esperado
Governança de IA	Como os documentos tratam ciclo de vida, responsabilização e monitoramento?	Enquadrar o problema como governança contínua, e não como validação pontual.
Cadeia de suprimento e proveniência	Quais componentes compõem o objeto de confiança?	Ampliar o foco do checkpoint para o stack sociotécnico e técnico-operacional.
Atualização e mudança	Quais evidências tornam uma promoção de versão defensável?	Definir critérios de admissão, pinagem, validação e aprovação formal.
Telemetria e observabilidade	Que fluxos de dados, metadados e chamadas externas precisam ser distinguidos?	Separar observabilidade interna de egress (saída externa) para terceiros.
Rollback e incidentes	Quais pré-condições tornam a reversão auditável?	Definir capacidade ensaiada de reversibilidade e resposta.
Regulação e jurisdição	Que fluxos acessórios podem	Mapear papéis, fluxos, bases normativas e suporte remoto.

Eixo analítico	Pergunta de extração	Resultado esperado
	reativar obrigações regulatórias?	

Fonte: elaborado pelos autores (2026), com base na matriz analítica aplicada ao corpus.

ESTADO DA ARTE E LACUNAS DE GOVERNANÇA OPERACIONAL

O estado da arte reúne recomendações robustas sobre governança de IA, desenvolvimento seguro, cadeia de suprimento, monitoramento e resposta a incidentes. Essas recomendações, porém, aparecem com frequência em tradições documentais separadas. Normas e frameworks tratam ciclo de vida e responsabilização; guias de segurança enfatizam implantação e incidentes; documentação técnica descreve opções de telemetria, modo offline e serialização; e a literatura científica discute riscos de privacidade, segurança, extração de modelos e cadeia de suprimento. O valor da revisão está justamente na leitura transversal dessas tradições, aproximando a formulação normativa da decisão operacional.

Essa dispersão cria uma lacuna prática para organizações que precisam decidir se um modelo generativo de terceiro pode entrar, permanecer, ser atualizado, monitorado ou retirado de um ambiente sensível. O problema, portanto, não é simplesmente a ausência de controles. Falta um padrão analítico que conecte a entrada do artefato, sua promoção de versão, seus fluxos informacionais e sua capacidade de reversão. Em ambientes sensíveis, não basta perguntar se o modelo “roda localmente” ou se a solução “parece segura”. A pergunta de governança deve ser outra: o artefato tem proveniência verificável, disciplina de saída mapeada e rollback executável?

A literatura sobre monitoramento pós-implantação reforça que métodos, terminologias e boas práticas ainda são dispersos, apesar do consenso sobre a necessidade de acompanhar sistemas de IA em produção (Rao et al., 2026). Esse ponto é relevante porque modelos generativos podem alterar comportamento em razão de dependências, prompts, conectores, dados de contexto e mudanças de versão. A governança, portanto, deve considerar não apenas a validação anterior à

entrada em produção, mas também evidências de funcionamento durante a operação e critérios para resposta.

Diante dessa lacuna, atualização, telemetria e rollback são tratados como dimensões indissociáveis. A primeira opera como mudança governada; a segunda, como disciplina de fluxos internos e externos; o terceiro, como capacidade material de desfazer uma decisão de implantação. A posição teórica assumida é que a governança de modelos generativos de terceiros executados localmente não se completa pela existência de política institucional, nem pela execução técnica isolada do modelo, mas pela articulação entre evidência de origem, controle de fluxos e reversibilidade material. O quadro 5 resume essa relação; a convergência dessas dimensões permite avaliar a admissão inicial do modelo, sua permanência, sua promoção de versão e sua retirada segura ao longo do ciclo de vida.

Quadro 5 - Relação entre estado da arte, lacuna e resposta conceitual

Dimensão	Achado recorrente no corpus	Lacuna operacional	Resposta proposta
Atualização	Normas e guias recomendam gestão de mudança, componentes e segurança de desenvolvimento.	A atualização de modelo ainda pode ser tratada como substituição técnica simples.	Promover versão somente com pinagem, validação, proveniência e aprovação formal.
Telemetria	Documentações técnicas e guias destacam logs, variáveis de ambiente, modo offline e monitoramento.	Observabilidade interna e egress (saída externa) para terceiros podem ser confundidos.	Separar logs internos, chamadas externas, retenção, minimização e revisão jurídica.
Rollback	Guias de implantação segura abordam resposta, recuperação e continuidade operacional.	Rollback pode existir apenas como intenção, sem versão íntegra ou ensaio.	Definir gatilhos, preservar versões estáveis, testar recuperação e registrar decisões.
Regulação	Regulação de dados e IA exige mapeamento de fluxos, papéis e obrigações.	Fluxos acessórios podem não ser percebidos como transferência ou tratamento relevante.	Mapear suporte, sincronização, telemetria e base normativa antes da operação.

Destarte, v.15, n.1, p. 68-90, jan./jun. 2026 DOI: 10.5281/zenodo.21285122

Fonte: elaborado pelos autores (2026), a partir da síntese da revisão.

FRAMEWORK PROPOSTO DE GOVERNANÇA

O quadro 6 sintetiza o framework proposto, organizado em quatro camadas para a governança de modelos generativos de terceiros executados localmente: admissão, implantação, operação e reversibilidade/resposta. Na admissão, define-se o que está sendo incorporado e de quem; na implantação, estabelece-se como o artefato entra no ambiente e com quais limites; na operação, acompanham-se monitoramento, registros internos e reavaliação periódica; e, na reversibilidade/resposta, determinam-se as condições para conter, reverter ou retirar versões em caso de desvio, vulnerabilidade ou evento operacional.

A lógica do framework é impedir que controles tardios compensem deficiências de entrada. Um bom registro de logs não elimina proveniência obscura; um plano de incidentes não substitui política de atualização; e a execução local não afasta, por si só, exposição informacional ou regulatória quando existem fluxos externos acessórios. Por isso, cada decisão deve estar associada a uma pergunta diretiva, uma evidência mínima e uma responsabilidade organizacional.

Quadro 6 - Arquitetura em quatro camadas para governança de modelos generativos de terceiros executados localmente

Camada	Pergunta diretiva	Entregáveis mínimos
Admissão	O que está sendo incorporado e de quem?	Dossiê do fornecedor; inventário de artefatos; registro de proveniência; licença; verificação de integridade; avaliação inicial de risco.
Implantação	Como o artefato entra no ambiente e com quais limites?	Plano de mudança; testes em ambiente segregado; dependências fixadas; política de rede; telemetria permitida; baseline

Camada	Pergunta diretiva	Entregáveis mínimos
		operacional.
Operação	Como o uso é monitorado e revisto?	Logs e métricas internos; revisão de chamadas externas; trilha de auditoria; revisão de saídas; reavaliação periódica de risco residual.
Reversibilidade e resposta	Como conter ou reverter um desvio?	Critérios de rollback; versões anteriores íntegras; hashes preservados; playbook; responsabilidades formais; recuperação ensaiada; revisão pós-incidente.

Fonte: elaborado pelos autores (2026), a partir da síntese conceitual da revisão documental.

Na camada de admissão, a análise concentra-se na identidade e na integridade do que será incorporado. A organização precisa registrar fornecedor, licença, versão, hash, origem, dependências declaradas e finalidade de uso. Também deve avaliar se o artefato foi obtido de fonte confiável, se sua documentação é suficiente, se há restrições contratuais ou regulatórias e se a finalidade pretendida é compatível com o risco residual.

Na implantação, o foco se desloca para a entrada controlada no ambiente. O modelo deve ser testado em espaço segregado, com dependências fixadas e política de rede explícita. A organização deve verificar se o modo offline funciona quando necessário, se chamadas externas foram bloqueadas ou justificadas, se logs não expõem dados sensíveis e se a configuração de runtime está documentada. Essa camada, portanto, funciona como portão técnico entre intenção de uso e operação real.

Durante a operação, a governança requer observabilidade interna, revisão de egress (saída externa) e reavaliação periódica. Logs e métricas devem apoiar

auditoria, detecção de desvio, investigação de incidentes e comparação com baseline. A retenção desses registros, contudo, precisa respeitar minimização, classificação de dados e segregação de acesso, pois logs podem conter prompts, trechos de documentos, respostas do modelo ou metadados sensíveis.

Na reversibilidade e resposta, o rollback deixa de ser uma restauração improvisada e passa a funcionar como requisito de admissibilidade. Um serviço só deve ser tratado como reversível quando houver versão estável preservada, integridade verificável, gatilhos formais, responsáveis definidos e ensaio periódico de recuperação. Sem isso, a reversão pode ser nominal, lenta ou insegura, justamente quando a organização precisa conter rapidamente um desvio ou vulnerabilidade.

MATRIZ DECISÓRIA, DISCUSSÃO E APLICAÇÃO ANALÍTICA

O quadro 7 traduz o framework em critérios mínimos de controle. Ela não cria controles novos nem substitui avaliação jurídica ou técnica especializada; organiza, em linguagem operacional, as condições sob as quais um modelo de terceiro pode entrar, permanecer, ser atualizado, monitorado ou retirado de ambiente corporativo sensível. A matriz também funciona como artefato de comunicação entre segurança da informação, arquitetura, operações, compras, jurídico e privacidade.

Quadro 7 - Matriz decisória sintética de riscos e controles mínimos

Dimensão	Risco principal	Controle e gatilho mínimo
Atualização	Introdução de artefato malicioso, vulnerável ou incompatível.	Promoção apenas após pinagem, validação em ambiente segregado, verificação de proveniência e aprovação formal de mudança.
Telemetria	Saída não mapeada de dados ou metadados para terceiros.	Operação condicionada a inventário de chamadas externas, separação entre observabilidade

Destarte, v.15, n.1, p. 68-90, jan./jun. 2026 DOI: 10.5281/zenodo.21285122

Dimensão	Risco principal	Controle e gatilho mínimo
		interna e telemetria de fornecedor, minimização de dados e revisão jurídica quando aplicável.
Rollback	Persistência de versão comprometida sem reversão rápida e auditável.	Serviço tratado como reversível apenas quando houver versão estável preservada, integridade verificável, gatilhos formais e ensaio periódico de recuperação.
Vazamento informacional	Exposição de prompts, logs, pesos, segredos operacionais ou dados memorizados.	Uso condicionado à classificação de dados, hardening de logs, segregação de acesso, limites de retenção e controle de extração.
Exposição jurisdicional	Fluxos transfronteiriços indevidos ou obrigações não mapeadas.	Integração condicionada ao mapeamento de fluxos, base normativa adequada e revisão de suporte, sincronização e serviços acessórios.

Fonte: elaborado pelos autores (2026), com base no framework de governança proposto.

Para demonstrar a aplicação analítica, pode-se considerar uma organização que pretende usar localmente um modelo generativo de terceiro para análise de documentos internos confidenciais. O exemplo tem caráter demonstrativo e não deve ser confundido com estudo de caso empírico; serve para mostrar como o

framework orienta a leitura operacional de uma decisão de governança. Na admissão, o modelo somente deveria avançar se houvesse hash registrado, licença revisada, dependências documentadas, origem verificável e avaliação inicial de risco. Na implantação, exigiria ambiente segregado, telemetria de fornecedor desativada quando aplicável, limitação explícita de acessos externos e baseline comportamental.

Durante a operação, o uso dependeria de logs e métricas internos, revisão periódica de chamadas externas, auditoria de acessos, revisão de saídas e reavaliação diante de novas dependências ou alterações de versão. Na etapa de reversibilidade, a organização precisaria definir gatilhos para retorno à última versão aceitável, preservando artefatos íntegros e evidências de decisão. O valor prático do framework está em converter perguntas vagas - como “o modelo funciona localmente?” ou “a solução parece segura?” - em verificações objetivas sobre artefatos, fluxos e reversibilidade.

Do ponto de vista de engenharia e computação, a proposta aproxima governança de práticas auditáveis de mudança, segurança, observabilidade e continuidade operacional. A decisão passa a incluir qual artefato foi admitido, quais dependências foram congeladas, quais fluxos externos são permitidos, quem aprovou a mudança, quais logs são mantidos, qual versão estável está preservada e qual procedimento foi ensaiado para reversão. Desse modo, o framework atua como ponte entre recomendações normativas e tarefas concretas de operação, preservando como etapa futura sua validação em contextos organizacionais reais.

Na prática, a matriz pode servir como roteiro de verificação em pilotos organizacionais, auditorias internas ou decisões de admissão de modelos de terceiros. Nesses usos, orienta a coleta de evidências, a identificação de responsabilidades e a definição de critérios mínimos de atualização, telemetria e rollback. Essa aplicação inicial reforça a utilidade gerencial do framework, sem apresentar o exemplo como validação empírica já concluída.

CONCLUSÕES

O artigo propôs um framework conceitual de governança para modelos generativos de terceiros executados localmente em ambientes corporativos de alto risco cibernético. A revisão bibliográfica e documental direcionada mostrou que

atualização, telemetria e rollback precisam ser examinados de forma integrada, pois a admissibilidade e a permanência de um modelo local dependem da avaliação conjunta da proveniência do artefato, da disciplina de saída e da capacidade real de reversão.

A principal contribuição é apresentar um critério analítico para admissão, promoção de versão, monitoramento e retirada segura de modelos generativos de terceiros. A aderência temática à Gestão, Informação e Tecnologia está na interface entre governança técnica, segurança de IA, cadeia de suprimento de software/modelos, observabilidade, resposta operacional e continuidade. Ao organizar evidências dispersas em arquitetura e matriz decisória, o artigo oferece uma base conceitual aplicável a ambientes em que falhas de configuração, fluxos externos ou ausência de rollback podem gerar impactos operacionais, informacionais e regulatórios.

Como limitação, o estudo não apresenta validação experimental, protótipo, comparação quantitativa de desempenho ou evidência empírica própria. A revisão documental é direcionada e funcional, de modo que não pretende exaurir toda a literatura sobre governança de IA ou segurança de LLMs. A suficiência do corpus foi definida pela reiteração funcional das categorias necessárias à proposta, e não por saturação estatística. Pesquisas futuras podem aplicar o framework em estudos de caso setoriais, propor indicadores de maturidade para atualização, telemetria e reversibilidade, e avaliar métricas de monitoramento pós-implantação em ambientes organizacionais aplicados. Essa agenda permite avançar da proposta conceitual para evidências práticas sobre sua utilidade, seus limites e suas condições de generalização.

Como síntese prática, a execução local deve ser tratada como decisão de engenharia e governança, e não apenas como preferência arquitetural. A organização pode reduzir dependência de provedores remotos, mas assume responsabilidades sobre proveniência, atualização, telemetria, registros, resposta e conformidade. Em relação a adoções pouco controladas, a diferença essencial está em transformar cada etapa de entrada, operação e retirada do modelo em evidência verificável.

REFERÊNCIAS

ARTIFICIAL INTELLIGENCE SAFETY AND SECURITY CENTER (AISC) et al. **Deploying AI systems securely: best practices for deploying secure and resilient AI systems.** [S.l.]: Joint Cybersecurity Information, 2024. Disponível em: https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF. Acesso em: 6 dez. 2025.

AUSTRALIAN CYBER SECURITY CENTRE (ACSC). **Artificial intelligence and machine learning:** supply chain risks and mitigations. Canberra: Australian Signals Directorate, 2026. Disponível em: https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/artificial-intelligence-and-machine-learning-supply-chain-risks-and-mitigations. Acesso em: 7 dez. 2025.

AUSTRALIAN CYBER SECURITY CENTRE (ACSC). **Engaging with artificial intelligence.** Canberra: Australian Signals Directorate, 2024. Disponível em: https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/engaging-with-artificial-intelligence. Acesso em: 13 dez. 2025.

AUTIO, Chloe et al. **Artificial Intelligence Risk Management Framework:** Generative Artificial Intelligence Profile. Gaithersburg: National Institute of Standards and Technology, 2024. Disponível em: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence. Acesso em: 14 dez. 2025.

AUTORIDADE NACIONAL DE PROTEÇÃO DE DADOS (ANPD). **Resolução CD/ANPD n. 19, de 23 de agosto de 2024**. Aprova o Regulamento de Transferência Internacional de Dados e o conteúdo das cláusulas-padrão contratuais. Brasília, DF: ANPD, 2024. Disponível em: https://www.gov.br/anpd/pt-br/acesso-a-informacao/institucional/atos-normativos/regulamentacoes_anpd/resolucao-cd-anpd-no-19-de-23-de-agosto-de-2024. Acesso em: 20 dez. 2025.

BATOOL, Amna; ZOWGHI, Didar; BANO, Muneera. AI governance: a systematic literature review. **AI and Ethics**, v. 5, p. 3265-3279, 2025. DOI: https://doi.org/10.1007/s43681-024-00653-w.

BOOTH, Harold et al. **Secure Software Development Practices for Generative AI and Dual-Use Foundation Models**: an SSDF community profile. Gaithersburg: National Institute of Standards and Technology, 2024. Disponível em: https://www.nist.gov/publications/secure-software-development-practices-generative-ai-and-dual-use-foundation-models-ssdf. Acesso em: 21 dez. 2025.

BOYENS, Jon et al. **Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations**. Gaithersburg: National Institute of Standards and Technology, 2024. DOI: https://doi.org/10.6028/NIST.SP.800-161r1-upd1. Disponível em: https://www.nist.gov/publications/cybersecurity-supply-chain-risk-management-practices-systems-and-organizations. Acesso em: 27 dez. 2025.

CARLINI, Nicholas et al. **Extracting Training Data from Diffusion Models**. arXiv, 2023. Disponível em: https://arxiv.org/abs/2301.13188. Acesso em: 28 dez. 2025.

HU, Qiang et al. **Large Language Model Supply Chain**: open problems from the security perspective. arXiv, 2024. Disponível em: https://arxiv.org/abs/2411.01604. Acesso em: 3 jan. 2026.
HUGGING FACE. **Environment variables.** Hugging Face Hub documentation. [S.l.], s.d.a. Disponível em: https://huggingface.co/docs/huggingface_hub/package_reference/environment_variables. Acesso em: 4 jan. 2026.

HUGGING FACE. **Installation.** Transformers documentation. [S.l.], s.d.b. Disponível em: https://huggingface.co/docs/transformers/installation. Acesso em: 10 jan. 2026.

ISO/IEC. **ISO/IEC 23894:2023:** Information technology - Artificial intelligence - Guidance on risk management. Geneva: International Organization for Standardization, 2023a. Disponível em: https://www.iso.org/standard/77304.html. Acesso em: 11 jan. 2026.

ISO/IEC. **ISO/IEC 42001:2023:** Information technology - Artificial intelligence - Management system. Geneva: International Organization for Standardization, 2023b. Disponível em: https://www.iso.org/standard/42001. Acesso em: 17 jan. 2026.

NATIONAL CYBER SECURITY CENTRE (NCSC) et al. **Guidelines for secure AI system development**. London: NCSC, 2023. Disponível em: https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf. Acesso em: 18 jan. 2026.

OWASP FOUNDATION. **OWASP Top 10 for Large Language Model Applications 2025.** [S.l.]: OWASP Foundation, 2025. Disponível em: https://owasp.org/www-project-top-10-for-large-language-model-applications/. Acesso em: 24 jan. 2026.

PYTORCH. Serialization semantics. **PyTorch documentation.** [S.l.], s.d. Disponível em: https://docs.pytorch.org/docs/stable/notes/serialization.html. Acesso em: 25 jan. 2026.

RAO, Anita et al. **Challenges to the Monitoring of Deployed AI Systems:** Center for AI Standards and Innovation. Gaithersburg: National Institute of Standards and Technology, 2026. DOI: https://doi.org/10.6028/NIST.AI.800-4. Disponível em: https://www.nist.gov/publications/challenges-monitoring-deployed-ai-systems-center-ai-standards-and-innovation. Acesso em: 31 jan. 2026.

SOUPPAYA, Murugiah; SCARFONE, Karen; DODSON, Donna. **Secure Software Development Framework (SSDF) Version 1.1:** recommendations for mitigating the risk of software vulnerabilities. Gaithersburg: National Institute of Standards and Technology, 2022. Disponível em: https://csrc.nist.gov/pubs/sp/800/218/final. Acesso em: 6 dez. 2025.

TABASSI, Elham. **Artificial Intelligence Risk Management Framework (AI RMF 1.0).** Gaithersburg: National Institute of Standards and Technology, 2023. DOI: https://doi.org/10.6028/NIST.AI.100-1. Disponível em: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10. Acesso em: 7 dez. 2025.

TRAMÈR, Florian et al. Stealing Machine Learning Models via Prediction APIs. In: **Usenix Security Symposium**, 25., 2016, Austin. Proceedings [...]. Berkeley: USENIX Association, 2016. Disponível em: https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/tramer. Acesso em: 13 dez. 2025.

UNIÃO EUROPEIA. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. **Official Journal of the European Union**, L, 2024/1689, 12 July 2024. Disponível em: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng. Acesso em: 14 dez. 2025.
WANG, Shenao et al. Large Language Model Supply Chain: a research agenda. arXiv, 2024. Disponível em: https://arxiv.org/abs/2404.12736. Acesso em: 20 dez. 2025.

YAO, Yifan et al. A survey on large language model (LLM) security and privacy: The Good, the Bad, and the Ugly. **High-Confidence Computing**, v. 4, n. 2, 2024. DOI: https://doi.org/10.1016/j.hcc.2024.100211.