xFabric
Unified Security Fabric Platform

WHITE PAPER TÉCNICO

AI Sec: governança de modelos e agentes

Inventário, avaliação contínua e evidência de conformidade para IA em produção

Superfície de exposição de aplicações LLM, controle de prompt e contexto, limites de agência e mapeamento de controles para OWASP LLM Top 10, NIST AI RMF e ISO/IEC 42001.

Documento técnico xFabric · revisão 2026.2 · uso corporativo

Contexto e fundamentação

Modelos em produção introduzem uma superfície nova — prompt, contexto recuperado, ferramenta invocada e dado de saída. Governança de IA é controle técnico com evidência, não política em documento.

Inventário antes de política

A primeira lacuna de qualquer programa de segurança de IA é o inventário. Modelos hospedados, endpoints de provedor externo, agentes com ferramentas, bases vetoriais e chaves de API se multiplicam por iniciativa de time, quase sempre fora do catálogo de aplicações.

O inventário precisa registrar propósito, dado de treino ou de contexto, provedor, jurisdição de processamento, ferramentas acessíveis pelo agente e responsável nomeado. Sem isso, qualquer controle subsequente cobre um subconjunto desconhecido da superfície.

A superfície específica de modelos generativos

As classes de exposição são distintas das de uma aplicação convencional e exigem controle próprio na fronteira de entrada e de saída.

- Injeção de prompt direta e indireta via conteúdo recuperado
- Vazamento de dado sensível por contexto recuperado sem filtro de autorização
- Uso excessivo de agência: ferramenta com permissão além do propósito declarado
- Envenenamento de base vetorial e de pipeline de ingestão
- Exfiltração por canal de saída não inspecionado

Avaliação contínua como controle, não como benchmark

Avaliação de modelo em segurança não mede qualidade de resposta; mede resistência a comportamento indesejado sob adversário. Conjuntos de avaliação devem cobrir injeção, evasão de política, alucinação com consequência operacional, viés em decisão automatizada e vazamento de contexto.

A execução precisa ser contínua e versionada com o modelo, o prompt de sistema e o conjunto de ferramentas. Uma alteração de qualquer um dos três invalida o resultado anterior e dispara nova avaliação antes da promoção.

Mapeamento para frameworks

Evidência de conformidade em IA se constrói por mapeamento entre controle técnico implementado e requisito de framework, com artefato coletado automaticamente a cada execução.

- OWASP Top 10 para aplicações LLM: cobertura por classe de exposição
- NIST AI RMF: funções de governar, mapear, medir e gerenciar
- ISO/IEC 42001: sistema de gestão de IA e registro de decisão
- Requisitos setoriais de explicabilidade em decisão automatizada

Ficha de inventário de modelo e agente

Nenhum controle de IA precede o inventário. Cada entrada registra, no mínimo, os campos abaixo, revisados a cada promoção de versão.

Campo	Descrição
Propósito	Caso de uso declarado e público atendido
Modelo e versão	Provedor, modelo, versão e prompt de sistema versionado
Dado de contexto	Fontes recuperadas e política de autorização aplicada
Ferramentas	Ações invocáveis pelo agente e permissão efetiva de cada uma
Jurisdição	Região de processamento e retenção do provedor
Responsável	Dono técnico nomeado e aprovador de negócio

Conjunto mínimo de avaliação adversária

- Injeção direta de prompt com tentativa de substituição de instrução
- Injeção indireta via documento recuperado e via conteúdo de terceiro
- Tentativa de extração de prompt de sistema e de segredo de configuração
- Vazamento de contexto pertencente a outro tenant ou a outro usuário
- Uso de ferramenta fora do propósito declarado
- Alucinação com consequência operacional em fluxo de decisão
- Viés em decisão automatizada sobre grupo protegido

Cada execução gera artefato assinado com resultado por classe, comparado ao limite acordado. Falha em classe crítica bloqueia a promoção.

Mapeamento de controles para frameworks

Framework	Requisito	Controle técnico
OWASP LLM Top 10	Injeção de prompt	Filtro de entrada, isolamento de contexto e validação de saída
OWASP LLM Top 10	Agência excessiva	Permissão efetiva por ferramenta e escopo de execução
NIST AI RMF	Medir	Avaliação contínua versionada com artefato retido
NIST AI RMF	Governar	Inventário com dono nomeado e aprovação de promoção

ISO/IEC 42001	Registro de decisão	Trilha imutável de prompt, contexto e resposta

Checklist de promoção para produção

- Entrada de inventário completa e aprovada
- Avaliação adversária executada na versão exata a promover
- Limite de agência revisado contra o propósito declarado
- Canal de saída inspecionado e política de dado sensível aplicada
- Plano de reversão e limite de consumo definidos
- Trilha de auditoria habilitada com retenção compatível com a jurisdição

Pontos de decisão

- Sem inventário de modelos, agentes e ferramentas, todo controle de IA cobre superfície desconhecida.
- Avaliação de segurança de modelo é controle de promoção, não relatório periódico.
- Agência de agente deve ser limitada por permissão efetiva, como qualquer outro princípio de identidade.