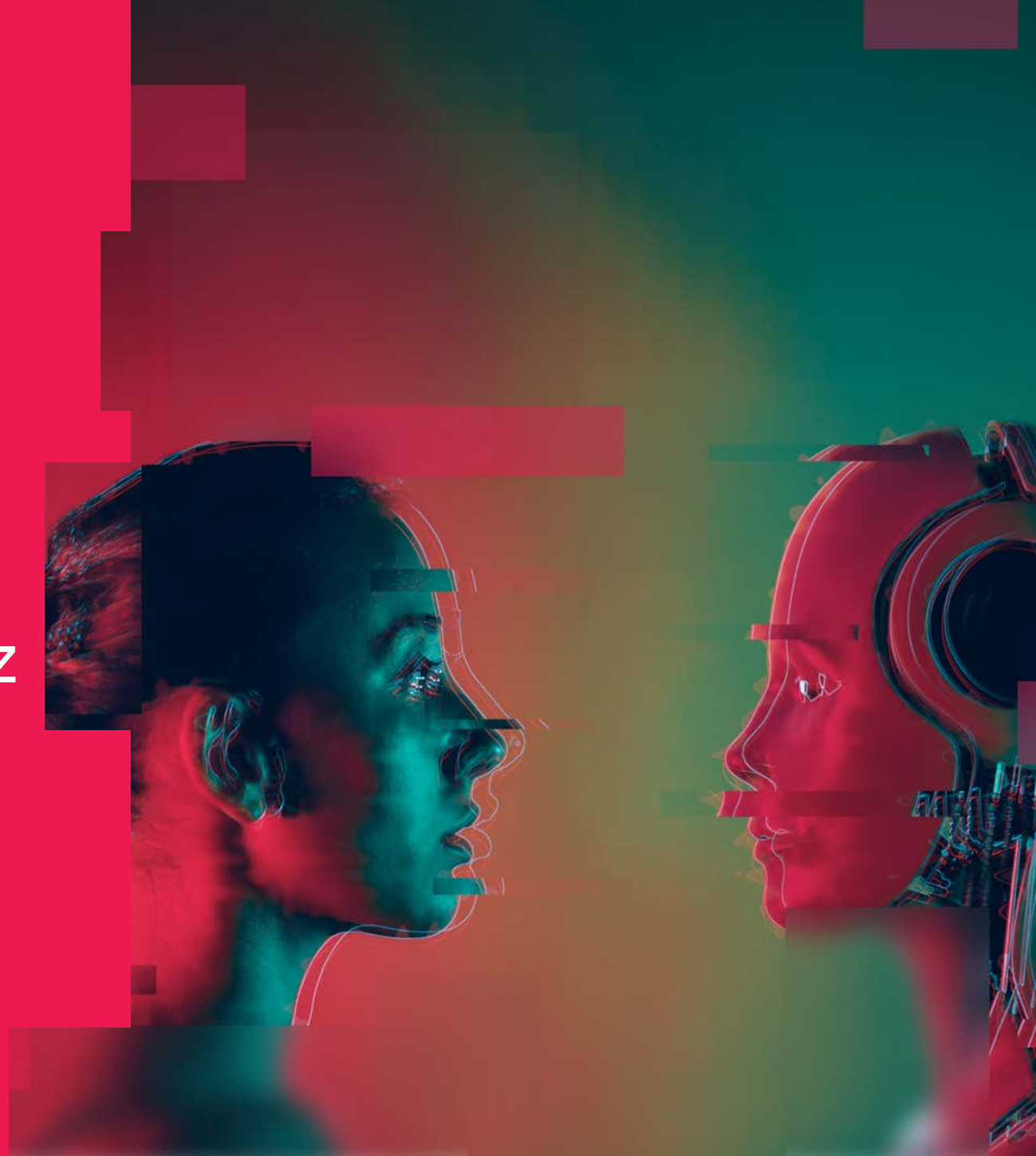
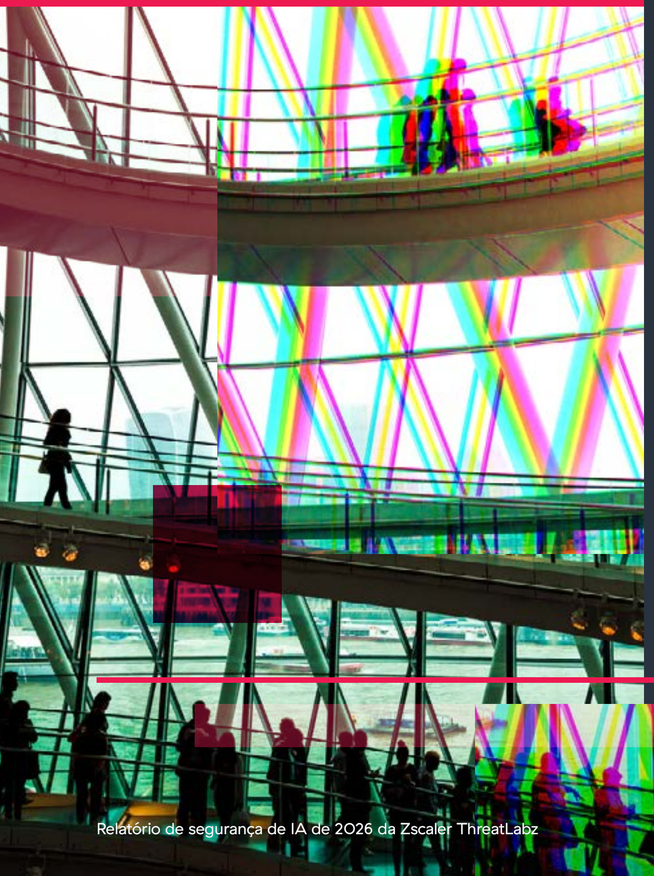
# Relatório de segurança de IA de 2026 da ThreatLabz



# Índice



# Executive Summary_

A realidade diária da IA em 2025 foi definida por velocidade, escala e movimento constante.

As empresas agora dependem de inteligência artificial e aprendizado de máquina (IA/ML) em toda a empresa para agir com mais rapidez, automatizar decisões e aumentar a produtividade. A IA auxilia o desenvolvimento, as comunicações, a pesquisa e as operações a um ritmo que teria parecido irreal há apenas alguns anos. Mas essa aceleração também trouxe consigo cada vez mais concessões: fluxos de dados mais sigilosos por meio de mais aplicativos de IA/ML, muitas vezes com menor visibilidade e menos proteções.

Essa crescente presença da IA ampliou a superfície de ataque das empresas, e os criminosos não perderam tempo em seguir esse exemplo no último ano. Barreiras mais baixas e maior realismo tornaram os ataques mais rápidos e convincentes, enquanto os primeiros indícios de uso indevido de agentes de IA e IA semiautônoma apontaram para uma mudança na forma como as ameaças estão evoluindo. Ao mesmo tempo, as organizações estão lidando com uma série crescente de riscos, desde IA paralela e embarcada até alucinações e modelos privados desprotegidos.

Como as empresas podem proteger ambientes onde a IA está presente em tudo, viabilizar a inovação orientada por IA e se defender contra ameaças baseadas em IA? (Tudo isso sem prejudicar o ritmo dos negócios, é claro).

O Relatório de segurança de IA de 2026 da Zscaler ThreatLabz explora como as empresas estão lidando com esse equilíbrio delicado. O relatório baseia-se na análise de 989,3 bilhões de transações de IA/ML observadas na Zscaler Zero Trust Exchange™ de janeiro a dezembro de 2025, proporcionando uma visão concreta de como a IA está sendo realmente usada (e restringida) em ambientes globais.

Os dados mostram uma aceleração contínua. A atividade de IA/ML empresarial aumentou 83.3% em relação ao ano anterior, enquanto os volumes de transferência de dados subiram 92,6%, atingindo mais de 18.000 terabytes (TB). Nessa escala, a IA se comporta menos como um conjunto de ferramentas isoladas e mais como uma infraestrutura sempre ativa, que movimenta e transforma continuamente os dados corporativos. O acesso, no entanto, está longe de ser irrestrito. As organizações bloquearam 39% das transações de IA/ML, refletindo preocupações persistentes em torno da exposição de dados, privacidade e aplicação de políticas.

Os padrões de uso também revelam onde o valor e o risco se cruzam. Os aplicativos de IA nos quais os funcionários mais confiam, como Codeium, Grammarly e ChatGPT, estão no centro da forma como o trabalho é realizado, promovendo os níveis mais altos de atividade e também aparecendo na vanguarda de nossas descobertas sobre riscos.

Em 2026, garantir a segurança da IA vai muito além do controle de aplicativos de IA/ML. Trata-se de garantir a segurança de como a IA é descoberta, construída, usada e governada em toda a empresa. As organizações precisam de visibilidade sobre o uso e os riscos da IA, proteções que fortaleçam os sistemas e dados de IA em tempo real e controles consistentes que protejam o acesso, mantendo a inovação em movimento. Este relatório explora as tendências e realidades que moldam a segurança de IA e fornece orientações para empresas que buscam reduzir riscos e  adotar a IA com segurança..

## O que isso significa para os líderes empresariais

- **A IA agora faz parte da infraestrutura empresarial.**
  Quase um trilhão de transações de IA sinalizam operações contínuas e ininterruptas. A IA deve ser governada com o mesmo rigor que a nuvem, a identidade e os dados para garantir uma adoção segura e escalável.

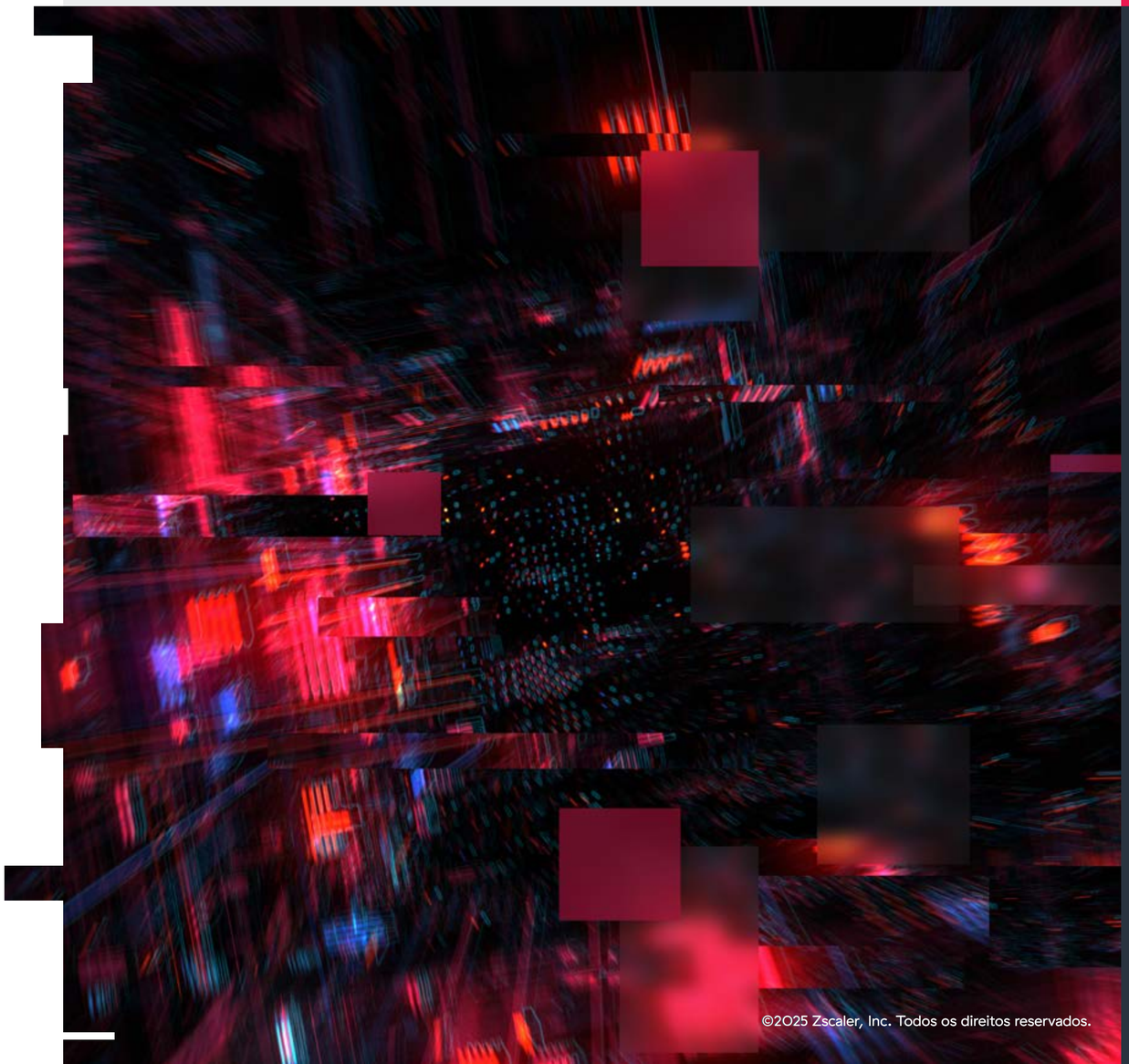
- **O risco de exposição de dados agora aumenta com o volume, não com a intenção.**
  A movimentação de dados em escala de petabytes por meio de fluxos de trabalho de IA aumenta a exposição devido à repetição e à velocidade, mesmo quando o uso é aprovado e está alinhado com a intenção do negócio.

- **A IA aprovada é a principal superfície de risco.**
  As ferramentas de IA convencionais e autorizadas respondem pela maioria das atividades de IA e interações de dados corporativas. Embora a IA paralela continue sendo uma preocupação fundamental, combater apenas as ferramentas não autorizadas não mitigará toda a extensão dos riscos e da exposição relacionados à IA.

- **A segurança está limitando a adoção da IA.**
  Com 39% das transações de IA bloqueadas, a aplicação de políticas está moldando ativamente a forma como a IA é utilizada. Isso reflete a governança em ação, e não a resistência à IA, à medida que os líderes buscam o equilíbrio entre a velocidade da inovação e a tolerância ao risco.

- **Os modelos de segurança tradicionais não são compatíveis com os fluxos de trabalho de IA.**
  Os controles projetados para atividades em ritmo humano e dados estáticos não conseguem acompanhar as interações de IA de alta frequência e automatizadas.

- **A vantagem competitiva favorecerá as organizações que conseguirem governar a IA em larga escala.**
  As empresas que permitirem o uso amplo da IA com controles robustos e integrados avançarão mais rapidamente do que aquelas forçadas a restringir totalmente o uso devido a riscos não gerenciados.



# Principais descobertas

A ThreatLabz analisou **989,3 bilhões de transações de IA e ML** na nuvem da Zscaler, de janeiro a dezembro de 2025. As principais descobertas a seguir são baseadas em dados abrangendo períodos de tempo variados* para análise comparativa.

**O uso de IA nas empresas continua sua forte trajetória ascendente.** A atividade de IA/ML aumentou 83% em relação ao ano anterior, atingindo quase um trilhão de transações em um ecossistema de mais de 3.400 aplicativos.

**As empresas enviam volumes cada vez maiores de dados para ferramentas de IA.** Um total de 18.033 TB de dados foi transferido para aplicativos de IA/ML, um aumento de 93% em relação ao ano anterior.

**Altas taxas de bloqueio indicam gerenciamento de riscos contínuo.** As empresas bloquearam 39% do total de transações de IA/ML, ressaltando as preocupações contínuas sobre a exposição de dados, a privacidade e o alinhamento de políticas à medida que o uso da IA se expande.

**A IA empresarial é extremamente vulnerável a ataques.** Os especialistas em testes de intrusão da Zscaler descobriram que a maioria dos sistemas de IA empresariais pode ser invadida em apenas 16 minutos e revelaram falhas críticas em 100% dos sistemas testados.

* Períodos de coleta de dados:

- Análise anual e comparativa com o mesmo período do ano anterior: janeiro a dezembro de 2025, com comparações anuais em relação ao mesmo período de 2024.
- Dados sobre violações de DLP e dados por país: junho a dezembro de 2025.

**A OpenAI domina como principal fornecedora de LLM.**
A OpenAI foi responsável pela grande maioria das transações empresariais baseadas em LLM (3 vezes mais que a Codeium), estabelecendo-se como o LLM padrão atual.

**O ChatGPT é responsável pela grande maioria das violações de DLP.** Em todos aplicativos de IA/ML analisados, o ChatGPT gerou 410 milhões de violações de políticas de prevenção contra perda de dados (DLP), confirmando os riscos empresariais associados a assistentes de IA de alto contexto.

**Aplicativos de produtividade integrados consolidam o uso de IA nas empresas.** O Grammarly se tornou o aplicativo n.º 1 por volume de transações, refletindo a dependência em IA que opera diretamente nos processos de comunicação e negócios.

**O setor de finanças e seguros e o setor de manufatura lideram novamente o uso de IA empresarial.** Pelo terceiro ano consecutivo, esses setores representaram a maior parcela do tráfego de IA/ML (23% e 20%, respectivamente) por trás de seus esforços de modernização e fluxos de trabalho de documentação complexos.

**Os Estados Unidos permaneceram a principal fonte de transações de IA/ML.** A atividade concentrou-se nos EUA, que representaram 38% das transações, seguidos pela Índia (14%) e Canadá (5%).

**A adoção da IA continua a expandir a superfície de ataque nas empresas.** O uso mais amplo da IA nos fluxos de trabalho corporativos criou mais caminhos para a exposição de dados e acessos, aumentando a probabilidade de vazamento de dados, uso indevido de prompts e ataques assistidos por IA, reforçando a necessidade de adotar uma arquitetura zero trust e controles de segurança baseados em IA.

# Tendências de uso_ de IA/ML

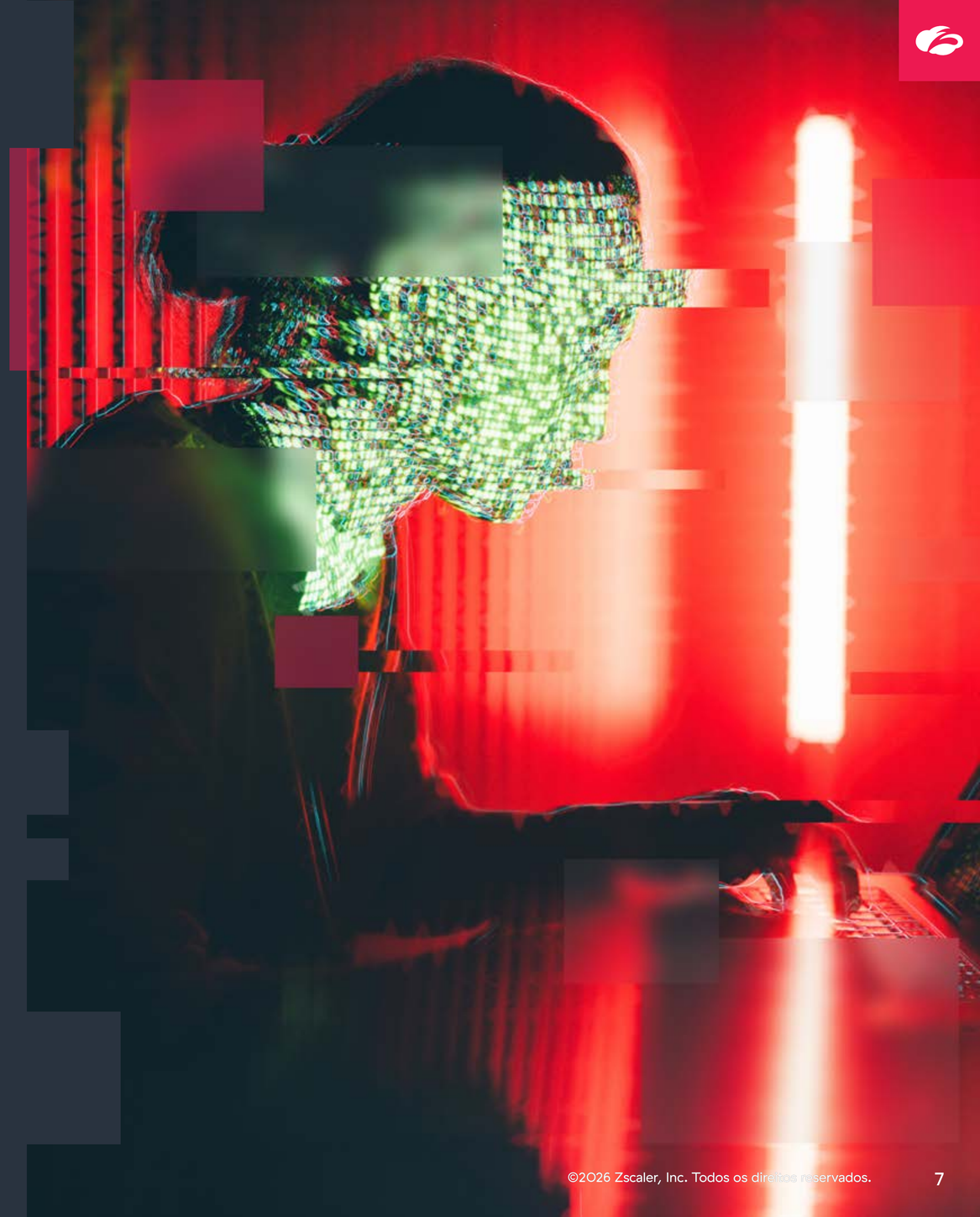
O uso empresarial da IA continuou sua ascensão acentuada e constante em 2025.

A análise da ThreatLabz sobre as tendências de uso de IA agora inclui mais de 3.400 aplicativos que promovem transações de IA/ML; quatro vezes mais do que no ano anterior. Embora muitos desses aplicativos gerem um tráfego limitado, o crescimento exponencial do próprio ecossistema de aplicativos é um indicador importante. Isso reflete a rapidez com que os recursos de IA estão se proliferando entre fornecedores, casos de uso e funções de negócios, expandindo tanto as oportunidades quanto a visibilidade.

Para entender como esse crescimento se traduz em uso empresarial no mundo real, a ThreatLabz analisou a atividade de IA/ML em várias camadas:

- **Transações gerais de IA/ML,** com base na categoria de URL, incluindo atividades permitidas e bloqueadas.
- **Classificação de fornecedores de LLM,** identificando quais provedores de modelos geram o maior tráfego de IA/ML e impulsionam os fluxos de trabalho de IA empresariais.
- **Principais aplicativos de IA/ML,** destacando os aplicativos específicos que promovem a atividade de IA empresarial e o volume de tráfego.
- **Uso departamental de IA,** com o mapeamento de aplicativos de IA de alto volume por departamento, para entender como a IA é utilizada no trabalho cotidiano.

Com essas perspectivas, pretendemos fornecer uma visão abrangente de como a IA está sendo efetivamente adotada nas empresas e onde o uso, a dependência e os riscos estão convergindo.



# Crescimento global em transações de IA/ML

As transações de IA/ML se aproximaram da marca de um trilhão em 2025, totalizando 989,3 bilhões. Grande parte desse crescimento está ligada a aplicativos de alto volume, como ChatGPT, Grammarly e Codeium.

## TENDÊNCIAS DE USO DE IA/ML POR VOLUME DE TRANSAÇÕES

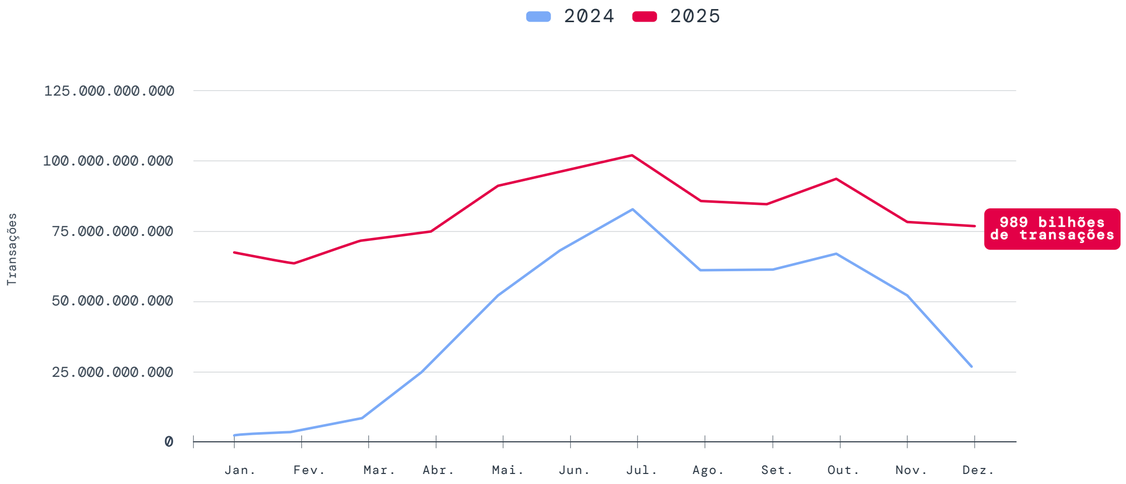



Figura 1: comparativo anual de transações de IA/ML (janeiro a dezembro de 2025)

PRINCIPAL DESCOBERTA

As atividades de IA/ML aumentaram 83% em relação ao ano anterior em um ecossistema de mais de 3.400 aplicativos.

Tal como nos anos anteriores, uma parte do tráfego enquadra-se na categoria de "aplicativos gerais de IA". Isso reflete transações de IA/ML que não correspondem a um aplicativo específico conhecido, mas são identificadas como relacionadas à IA pela categorização de URLs baseada em IA/ML da Zscaler, que analisa textos, imagens e outros sinais de conteúdo para reconhecer atividades relacionadas à IA. Novos aplicativos de IA surgem mais rapidamente do que podem ser classificados manualmente, tornando essencial detectar fontes de tráfego de IA anteriormente desconhecidas e submetê-las à aplicação de políticas de segurança.

Salvo indicação em contrário, as análises subsequentes neste relatório focaram-se exclusivamente em aplicativos classificados. Essa abordagem nos oferece visibilidade sobre a adoção da IA por meio de aplicativos de IA/ML já estabelecidos.

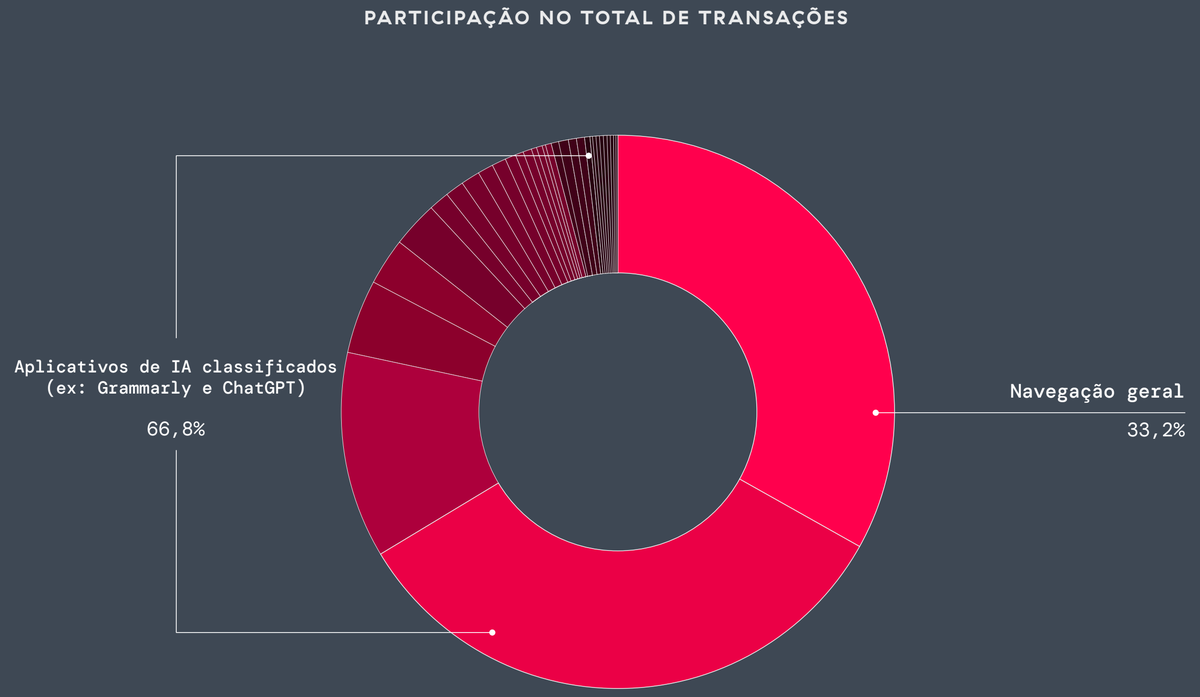



Figura 2: distribuição de transações de IA/ML em aplicativos de IA gerais e classificados

# Principais fornecedores, aplicativos e departamentos de LLM

Analisar o uso de IA empresarial por meio de fornecedores de LLM oferece uma visão única de como a IA está operando em grande escala. Embora os funcionários interajam diariamente com aplicativos e recursos individuais, os padrões de transação mostram quais provedores de modelo estão consistentemente por trás dessas experiências. A visibilidade em nível de fornecedor ajuda a compreender como a adoção de IA está evoluindo de forma menos aparente.

## Principais conclusões dos fornecedores de LLM

- **A OpenAI** liderou com folga o mercado de fornecedores de LLM em 2025, contabilizando 131 bilhões de transações, mais de três vezes o volume de seu concorrente mais próximo. O lançamento do GPT-5, em agosto, ampliou sua adoção em áreas como programação, raciocínio multimodal e execução de tarefas complexas. As opções expandidas de API empresarial da OpenAI, incluindo maior privacidade e isolamento de modelos, também reforçaram seu papel como plataforma para Copilots e recursos de SaaS com IA integrada.

- **A Codeium** (rebatizada como Windsurf em 2025) emergiu como a segunda maior fonte de tráfego corporativo de LLM (42 bilhões de transações). A adoção provavelmente foi impulsionada por seus modelos proprietários focados em programação, que aparecem frequentemente em fluxos de trabalho de desenvolvimento de software e ambientes de engenharia. Esse resultado espelha a análise departamental a seguir, em que a engenharia aparece como o departamento mais ativo no uso de IA.

- **A Perplexity** ficou em terceiro lugar em volume de transações no ano passado (12 bilhões de transações). Além da busca com inteligência artificial, ela também opera LLMs proprietários que alimentam seu mecanismo de respostas. Consequentemente, o uso corporativo reflete uma crescente dependência de pesquisa e síntese de conhecimento assistidas por IA.

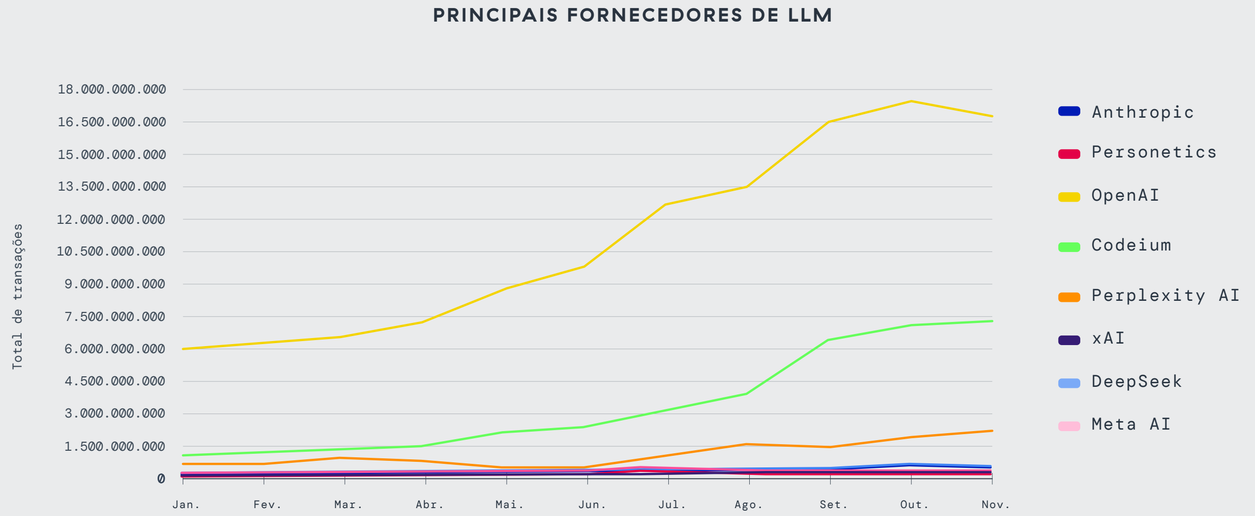



Figura 3: tendências de transações de fornecedores de LLM ao longo de 2025

O volume de transações permanece altamente concentrado em um conjunto de aplicativos amplamente adotados que se integram diretamente a fluxos de trabalho: pesquisa, edição, redação, programação, tradução e colaboração.

## Principais descobertas sobre aplicativos

- **O Grammarly** surgiu como o aplicativo de IA/ML mais ativo em ambientes corporativos (38,7% do total de transações), ultrapassando o ChatGPT em volume total de transações. Com recursos que vão desde resumos até reescrita avançada e mudanças de tom, fica fácil entender por que o Grammarly se destaca nos fluxos de trabalho de conteúdo corporativo do dia a dia.
- **O ChatGPT** continuou sendo um assistente de uso geral dominante (14,2%), amplamente utilizado em diversas funções, como pesquisa, redação e análise, tornando-se um ponto de contato comum para dados corporativos.
- **O Codeium** entrou no top cinco (5%), mostrando como a IA se tornou parte integrante do trabalho de desenvolvimento de software, onde código-fonte e lógica proprietária são processados rotineiramente.
- **O DeepL** continuou a apresentar forte adoção em organizações globais (3,3%), auxiliando na comunicação multilíngue em conteúdo essencial para os negócios.
- **O Microsoft Copilot** completou os cinco primeiros (3%), impulsionado por sua profunda integração ao Microsoft 365 e por seu papel na automatização de tarefas diárias de produtividade.

PRINCIPAIS APLICATIVOS DE IA

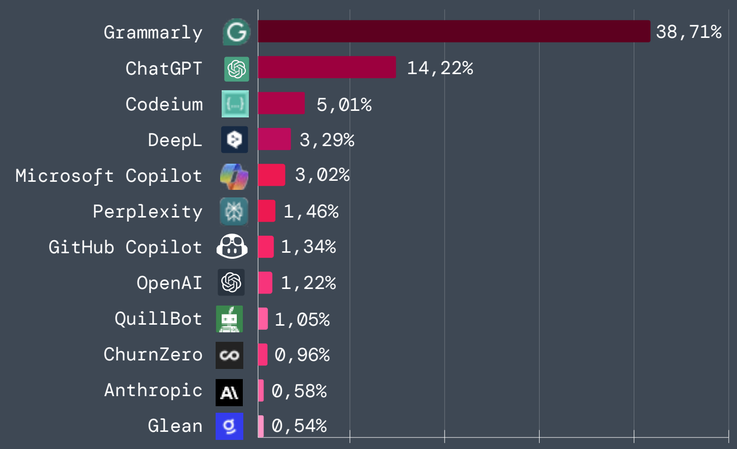



Figura 4: percentual do total de transações de IA/ML geradas pelos principais aplicativos de IA

Observação: a Zscaler Zero Trust Exchange rastreia transações do ChatGPT independentemente de outras transações da OpenAI em geral.

20 PRINCIPAIS APLICATIVOS DE IA/ML POR VOLUME DE TRANSAÇÕES

| Aplicativos | Total de transações |
|---|---|
| Grammarly | 327.311.080.013 |
| ChatGPT | 120.227.890.252 |
| Codeium | 42.337.652.986 |
| DeepL | 27.847.680.087 |
| Microsoft Copilot | 25.503.137.940 |
| Perplexity | 12.386.054.978 |
| GitHub Copilot | 11.348.420.722 |
| OpenAI | 10.352.420.115 |
| QuillBot | 8.913.115.535 |
| ChurnZero | 8.153.526.358 |
| Anthropic | 4.922.983.385 |
| Glean | 4.542.501.122 |
| GliaCloud | 3.249.239.347 |
| Claude | 2.850.954.278 |
| Google Gemini | 2.604.461.019 |
| SundaySky | 2.483.835.170 |
| Yellow Messenger | 1.734.555.650 |
| Cresta | 1.585.454.178 |
| Poe | 1.483.703.558 |

Além de identificar quais aplicativos de IA dominam o uso geral, o próximo nível de análise desloca o foco das ferramentas para as equipes.

A ThreatLabz mapeou o tráfego de IA/ML em um conjunto definido de departamentos comuns da empresa para melhor compreender como a IA está sendo usada na prática. Essa perspectiva concentra-se nos aplicativos com uso substancial (pelo menos um milhão de transações) e associa-os ao departamento em que são mais frequentemente utilizados. As porcentagens apresentadas refletem o uso relativo dentro desse conjunto específico de departamentos e aplicativos, e não o tráfego total de IA da empresa.

## Principais conclusões por departamento

- **O setor de engenharia** liderou o uso de IA nas empresas, representando 48,9% das transações de IA/ML dentro dessa visão delimitada. As equipes de engenharia, em particular, integram IA aos ciclos diários de desenvolvimento, onde mesmo pequenos ganhos de eficiência se acumulam rapidamente ao longo das versões.

- **A área de TI** seguiu de perto como um setor dependente de IA, representando 31,8% das atividades. O uso de IA em TI tende a apoiar a eficiência operacional, incluindo suporte ao sistema, solução de problemas e automação de processos internos.

- **O marketing** ficou em terceiro lugar no uso de IA empresarial (6.9%) desta análise. A adoção no marketing está mais distribuída entre fluxos de trabalho orientados a conteúdo e fluxos de trabalho focados em design, resultando em volumes de transação gerais estáveis, porém menores, em comparação com os departamentos técnicos.

### PARTICIPAÇÃO DE TRANSAÇÕES POR DEPARTAMENTO

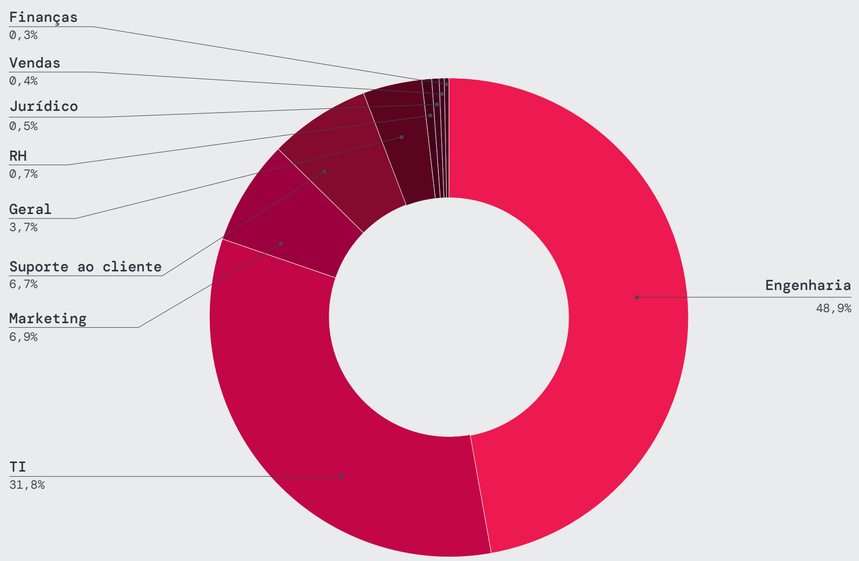



Figura 5: participação em transações de IA/ML por departamentos centrais da empresa

# Transações bloqueadas

Em 2025, as organizações também intensificaram o controle sobre a IA empresarial. Preocupações com a exposição de dados, privacidade e conformidade as levaram a bloquear 39,2% do total de transações de IA/ML, reforçando a governança de IA como parte padrão das operações diárias de segurança.

Os aplicativos mais impactados pela aplicação de controles também estavam entre os aplicativos de IA mais utilizados nas empresas. O Grammarly representou a maior parcela individual de atividades bloqueadas: 171,2 bilhões de transações bloqueadas, o que corresponde a 44,2% de todas as transações de IA/ML bloqueadas. Os aplicativos de IA de uso amplo também permaneceram sob análise. O ChatGPT e o Microsoft Copilot foram bloqueados com frequência, registrando 5,7 bilhões e 4,1 bilhões de transações bloqueadas, respectivamente, à medida que o acesso a dados não estruturados continua a aumentar o risco de compartilhamento não intencional de informações corporativas sigilosas.

Assistentes de programação com IA, incluindo Codeium e Tabnine, também eram frequentemente bloqueados para limitar a exposição de código proprietário e artefatos de desenvolvimento. Ferramentas de transformação de linguagem e conteúdo, como QuillBot e DeepL, enfrentaram controles semelhantes, refletindo esforços mais amplos para limitar o compartilhamento de conteúdo com modelos externos.

### PRINCIPAIS APLICATIVOS DE IA BLOQUEADOS

| | |
|---|---|
| 1 | Grammarly |
| 2 | GitHub Copilot |
| 3 | ChatGPT |
| 4 | Microsoft Copilot |
| 5 | QuillBot |
| 6 | Codeium |
| 7 | DeepL |
| 8 | Tabnine |
| 9 | Poe |
| 10 | Perplexity |

# Dados transferidos para aplicativos de IA

O volume de transações por si só não captura completamente como as empresas estão usando a IA. Para contextualizar, a ThreatLabz também examinou a quantidade de dados transferidos entre ambientes corporativos e aplicativos de IA/ML.

Ao longo do último ano, a transferência de dados empresariais para aplicativos de IA/ML continuou a aumentar, atingindo 18.033 terabytes (TB); um aumento de 93% em relação ao ano anterior. Um subconjunto de aplicativos populares amplamente adotados foi responsável pela maior parte dessa movimentação de dados. O Grammarly manteve-se como o aplicativo líder nesse quesito, com 3.615 TB de dados transferidos. Logo em seguida veio o ChatGPT (2.021 TB), seguido pela OpenAI (865 TB), DeepL (625 TB) e Codeium (387 TB); aplicativos que abrangem casos de uso que normalmente lidam com dados empresariais de alto valor.

À medida que a IA se torna mais integrada ao trabalho diário, mais dados corporativos transitam por ela. Analisar tanto o tráfego quanto o volume de dados ajuda a identificar onde o uso de IA está se expandindo e onde a segurança e a supervisão são mais importantes.

**PARTICIPAÇÃO DE DE DADOS TRANSFERIDOS**

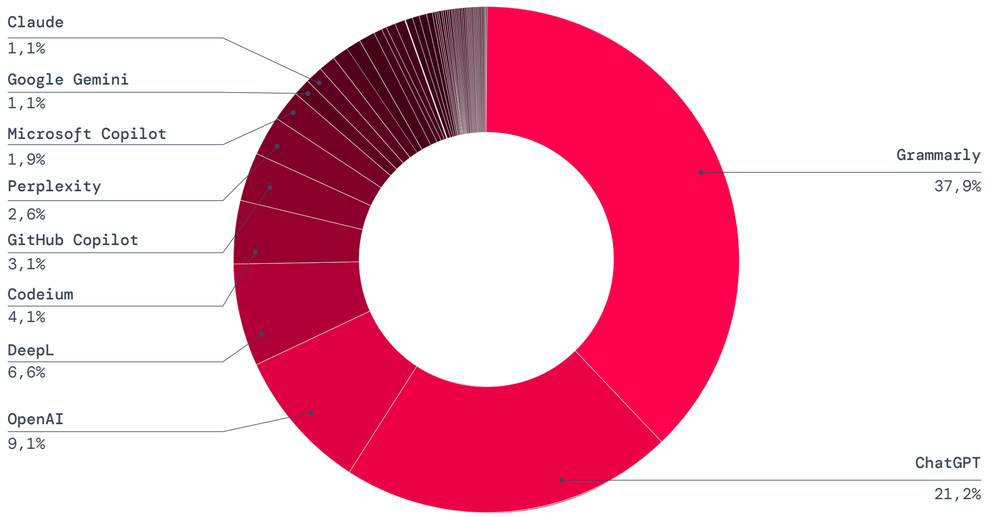
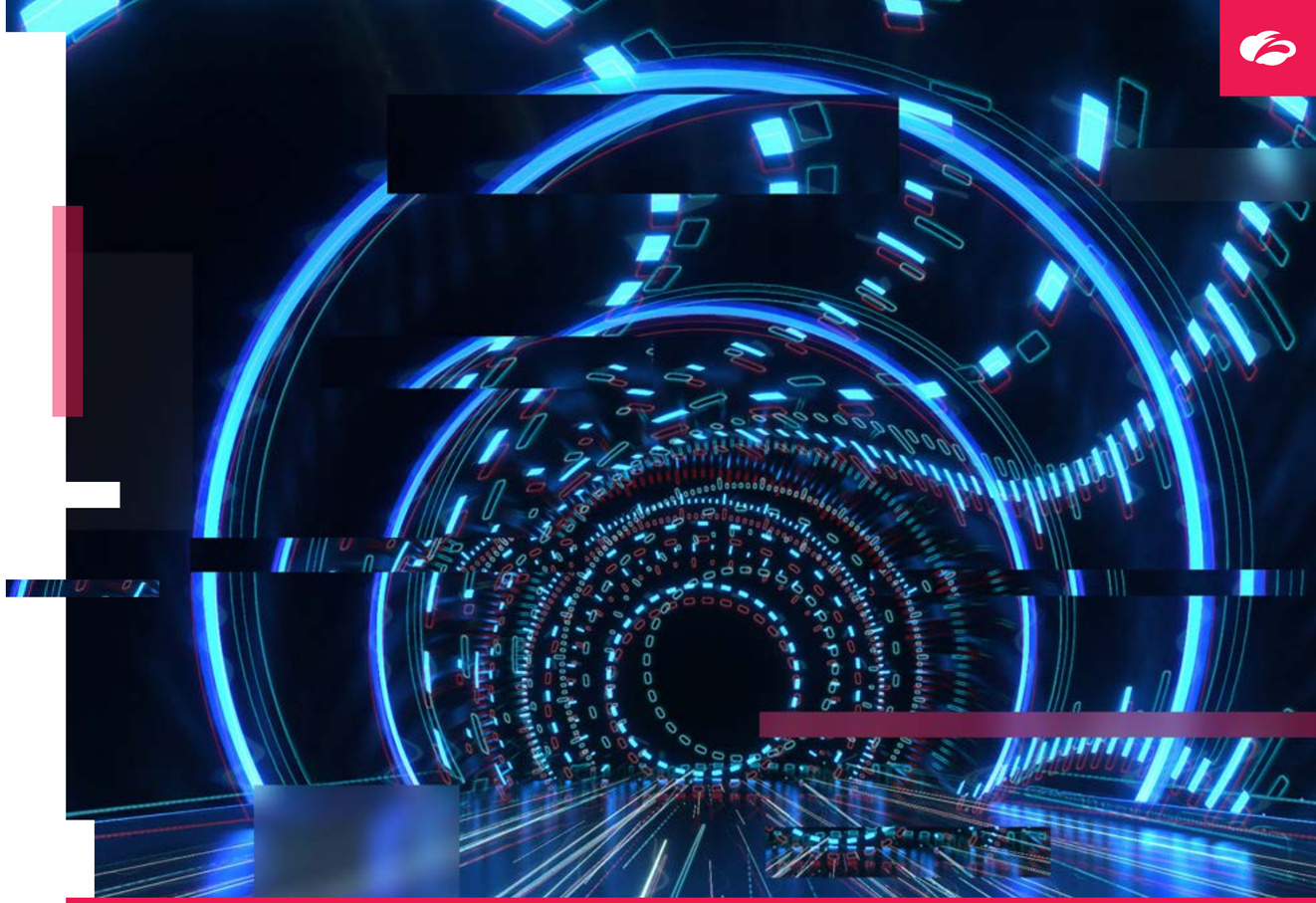



Figura 6: principais aplicativos de IA/ML pela porcentagem do total de dados transferidos



**PRINCIPAL DESCOBERTA**

Um total de **18.033 TB de dados foram transferidos para aplicativos de IA/ML**; um aumento de 93% em relação ao ano anterior.

# Perda de dados para aplicativos de IA

A capacidade da IA de acelerar o trabalho, indo da ideia à conclusão em minutos, tem um preço alto: dados sigilosos podem ser compartilhados com modelos externos em *segundos*. Além disso, com recursos de IA integrados em aplicativos e serviços de SaaS comuns, o conteúdo geralmente é transmitido automaticamente, aumentando a probabilidade de exposição despercebida.

**Prevenir a perda de dados para modelos externos tornou-se uma das prioridades de segurança mais importantes do ano.**

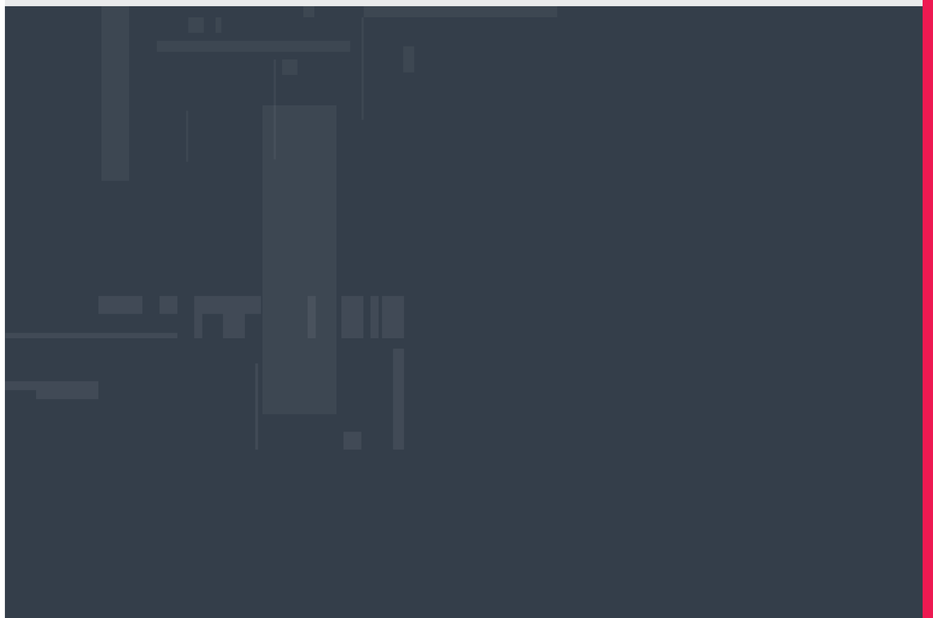
Na nuvem da Zscaler, as violações das políticas de DLP relacionadas à IA continuam sendo um dos sinais mais claros desse risco crescente. Essas violações ocorrem quando informações sigilosas, como registros financeiros, informações de identificação pessoal (PII), código-fonte, dados de saúde e outros conteúdos regulamentados, tentam sair da organização por meio de um aplicativo de IA e são bloqueadas pelas políticas da empresa. Sem a solução de DLP com IA da Zscaler, esses dados teriam ficado expostos a modelos de terceiros fora do controle da empresa.

Os aplicativos de IA mais arriscados tendem a ser aqueles que os funcionários usam sem pensar: assistentes de escrita, ferramentas de programação ou recursos de IA integrados a softwares de colaboração. Sua conveniência é exatamente o que os torna de maior risco; eles veem o mesmo conteúdo sigiloso que os funcionários, muitas vezes no momento em que ele é criado.

As tendências de violação mostram que as interações com IA envolvem, na maioria das vezes, alguns dos dados mais sigilosos da empresa.

APLICATIVOS DE IA/ML COM MAIS VIOLAÇÕES DE POLÍTICA DE DLP

| Aplicativos | Contagem de violações de DLP |
|---|---|
| ChatGPT | 410.181.006 |
| Codeium | 242.263.311 |
| GitHub Copilot | 31.223.009 |
| Claude | 14.417.246 |
| Wordtune | 5.161.758 |
| DeepL | 2.037.613 |
| QuillBot | 1.960.391 |
| Microsoft Copilot | 1.858.952 |
| Perplexity | 1.235.129 |
| Google Gemini | 841.374 |



**As violações de DLP do ChatGPT aumentaram 99,3%** em relação ao ano anterior. As violações mais comuns específicas do ChatGPT incluíram vazamento de nomes e identificadores nacionais; possivelmente registros de clientes ou dados de identidade.

As violações de DLP corporativa relacionadas ao **Codeium aumentaram 100%** em relação ao ano anterior, sugerindo um risco maior de vazamento de código-fonte e lógica proprietária.

O que mais se destaca nas principais violações de DLP com IA é o alcance global da exposição. Identificadores nacionais, dados de pagamento, código-fonte e informações médicas, cada um regido por regulamentações regionais rigorosas, estão surgindo cada vez mais nas interações de IA.

### AS 10 PRINCIPAIS VIOLAÇÕES DE POLÍTICAS DE DLP COM IA

| | |
|---|---|
| 1 | Vazamento de nome |
| 2 | Número de segurança social (EUA) |
| 3 | Número da empresa (Japão) |
| 4 | Número do serviço nacional de saúde (Reino Unido) |
| 5 | Código-fonte |
| 6 | Número do Medicare (Austrália) |
| 7 | Número de identificação nacional de fornecedor (EUA) |
| 8 | Número do seguro social (Canadá) |
| 9 | Medical information |
| 10 | Informações sobre cartões de crédito |

Essas tendências de DLP correspondem à mesma dinâmica de falhas observada quando sistemas de IA são testados em condições reais de adversidade: ocorrem falhas críticas, frequentemente por meio de interações comuns em vez de ataques sofisticados. Saiba mais em **O que realmente está falhando nos sistemas de IA empresariais**, abaixo.

Para saber como mitigar a perda de dados em aplicativos de GenAI, leia **Como as empresas estão implementando a GenAI com segurança**, abaixo.

# A ascensão da IA embarcada

Nem todo o uso de IA empresarial se manifesta em ferramentas de IA generativa independentes. Cada vez mais, isso acontece por meio da IA embarcada: recursos incorporados em aplicativos do dia a dia que não são classificados como aplicativos de GenAI, como resumos, recomendações ou insights automatizados que acionam a IA apenas em determinados momentos. Essas funcionalidades geralmente parecem atualizações naturais e esperadas para ferramentas que os usuários já utilizam. É também por isso que é fácil ignorar o fato de que a IA embarcada também interage com os dados corporativos sem a mesma visibilidade ou proteções que os aplicativos de IA independentes, tornando-se uma dimensão mais discreta, porém cada vez mais importante, da segurança na adoção da IA. Consequentemente, a IA embarcada representa uma das fontes de risco de IA empresarial que cresce mais rapidamente e é uma das menos visíveis.

Essa mudança de categoria é importante porque a IA embarcada foi projetada para aumentar a produtividade, incorporando mais contexto. O mesmo princípio de projeto também pode aumentar a exposição se a governança e os controles não acompanharem o ritmo. Os seguintes padrões de ameaças são comumente associados a recursos de IA embarcada em aplicativos empresariais.

## Principais observações

### COMPARTILHAMENTO EXCESSIVO DECORRENTE DE PERMISSÕES HERDADAS

A IA embarcada normalmente depende de controles de acesso e permissões de conteúdo já existentes. Se uma organização tiver acesso amplo por padrão, assinaturas de grupo desatualizadas ou espaços de colaboração excessivamente compartilhados, a IA embarcada pode, involuntariamente, expor informações sigilosas a usuários que tecnicamente têm acesso, mas não precisam dessas informações para desempenhar suas funções. Na prática, isso pode transformar uma proliferação de permissões já existente em uma exposição de dados mais rápida e mais visível.

### MANIPULAÇÃO INDIRETA DE PROMPTS POR MEIO DE CONTEÚDO EMPRESARIAL

A IA embarcada costuma ler conteúdo corporativo, como e-mails, incidentes, documentação, registros de bate-papo e anexos, como parte de suas operações normais. Isso introduz um risco, pois instruções ocultas ou conteúdo adversário podem influenciar a forma como a IA responde, o que ela prioriza ou como apresenta as informações. Quando os recursos de IA são estreitamente integrados aos fluxos de trabalho, o próprio conteúdo pode se tornar um canal de manipulação.

### EXPOSIÇÃO NA CADEIA DE SUPRIMENTOS DE MODELOS E CONECTORES

As funcionalidades de IA embarcada frequentemente dependem de múltiplos componentes. Isso pode incluir provedores de modelos, camadas de recuperação que extraem conteúdo de sistemas corporativos e conectores que se integram em aplicativos SaaS e repositórios de dados. Cada componente pode introduzir novos limites de confiança e novos vetores de mudança. À medida que as funcionalidades evoluem, o perfil de risco pode mudar devido a atualizações, alterações de configuração ou novas integrações ativadas.

### RISCOS DE AÇÃO E AUTOMAÇÃO EM FLUXOS DE TRABALHO HABILITADOS POR IA

À medida que as funcionalidades da IA vão além do resumo e da elaboração de rascunhos, passando a executar tarefas, a superfície de risco aumenta. Se um recurso de IA puder desencadear ações, recomendar alterações, gerar código ou preencher registros, erros ou resultados manipulados poderão se tornar problemas operacionais. Mesmo sem a execução direta de ações, os resultados gerados por IA podem influenciar decisões e fluxos de trabalho subsequentes de maneiras difíceis de auditar.

### EXPLORAÇÕES REAIS ENVOLVENDO IA EMBARCADA PERMITEM A FÁCIL EXFILTRAÇÃO DE DADOS

Dois exemplos de exploração amplamente divulgados no ecossistema do Copilot ilustram como a baixa interação do usuário ainda pode resultar em alto risco de IA embarcada:

- **O EchoLeak** é descrito como uma vulnerabilidade de injeção de prompt sem cliques no Microsoft 365 Copilot, que poderia permitir a exfiltração de dados por meio de padrões normais de assimilação de e-mails.
- **O Reprompt** é um ataque de clique único que utilizava prompts personalizados por meio de parâmetros de URL para desencadear comportamentos indesejados e vazamento de dados.

Olhando para o futuro, à medida que mais fornecedores de SaaS oferecem IA por padrão e expandem os recursos embarcados, as empresas precisarão estender a visibilidade, a governança e a proteção de dados da IA para os aplicativos e fluxos de trabalho onde a IA opera implicitamente.

# Uso de IA/ML por setor

A adoção da IA acelerou em todos os setores em 2025, com todos os segmentos abrangidos pela nuvem da Zscaler apresentando aumentos ano a ano em atividades de IA/ML. Mas o ritmo e o grau de maturidade da adoção variam bastante. Em alguns setores, ela já está apresentando resultados concretos. Em outros casos, ainda está encontrando seu lugar.

As empresas de **finanças e seguros** representam a maior parte (23,3%) do tráfego de IA/ML pelo segundo ano consecutivo. Bancos e seguradoras são naturalmente os primeiros a adotar a IA, dada a forte dependência de suas operações em relação a dados, análises e automação. A **manufatura** manteve a segunda posição, com 19,5% do total de transações de IA/ML, o que pode ser atribuído ao seu investimento em automação orientada por IA, controle de qualidade, otimização da cadeia de suprimentos, entre outros. Os setores de **tecnologia e comunicação** e **educação** registraram os maiores aumentos anuais, conforme destacado abaixo.

PARTICIPAÇÃO DE TRANSAÇÕES DE IA POR SETOR VERTICAL

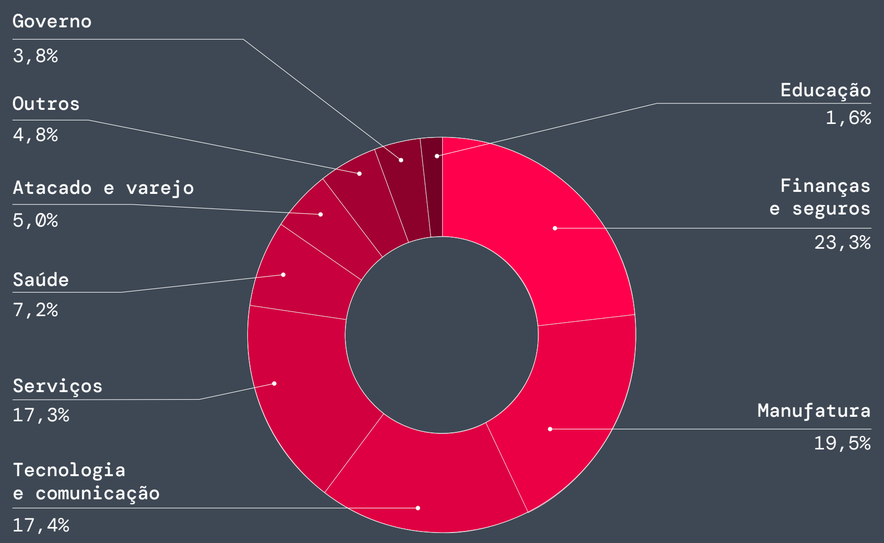



Figura 7: setores que promovem as maiores proporções de transações de IA

PARTICIPAÇÃO DE TRANSAÇÕES DE IA BLOQUEADAS POR VERTICAL

| Vertical | % de transações de IA bloqueadas |
|---|---|
| Finanças e seguros | 39,1% |
| Manufatura | 22,1% |
| Serviços | 13,5% |
| Saúde | 8,5% |
| Tecnologia e comunicação | 6,8% |
| Governo | 4,0% |
| Outros | 3,4% |
| Atacado e varejo | 2,0% |
| Educação | 0,6% |

O uso da IA não ocorre isoladamente; ele é influenciado pelos riscos específicos do setor, pelas expectativas de conformidade e pelo nível de evolução dos programas de segurança.

Padrões em transações de IA/ML bloqueadas revelam como os setores estão equilibrando de maneiras diferentes a adoção da IA com o gerenciamento de riscos. O setor de finanças e seguros não apenas gerou a maior parte da atividade de IA, como também bloqueou cerca de 40% dessas transações. A alta taxa de bloqueios reflete mais do que cautela: é a realidade de operar em um ambiente altamente regulamentado, onde se esperam controles mais rígidos sobre o uso de IA.

O setor manufatureiro, o segundo mais ativo em volume de transações de IA, bloqueou aproximadamente 22% do seu tráfego de IA. Isso sugere um meio-termo pragmático, visto que os fabricantes implementam IA amplamente, mas ainda aplicam uma supervisão significativa para evitar o uso indevido e proteger contra vazamento de dados; especialmente em ambientes de IoT/OT.

DESTAQUES DO SETOR

# O setor de finanças e seguros permanece como o mais impulsionado por IA, com 230 bilhões de transações

O setor de finanças e seguros foi o maior gerador de atividades de IA na nuvem da Zscaler em IA/ML, representando quase um quarto de todo o uso corporativo. Grande parte desse volume vem de ferramentas de produtividade do dia a dia. Grammarly, ChatGPT e Microsoft Copilot foram os aplicativos de IA mais usados em bancos e seguradoras pelo segundo ano consecutivo. Equipes em diversas organizações usam essas ferramentas para resumir pesquisas, lidar com documentação de conformidade, detectar fraudes, agilizar sinistros, auxiliar na contratação de seguros e executar outras tarefas essenciais. Essas tendências refletiram o crescimento do setor como um todo. De acordo com a pesquisa AI Adopter 2025, do Morgan Stanley[1], a adoção de IA no setor de seguros saltou de 48% para 71% em meados do ano, e de 66% para 73% em empresas de serviços financeiros.

A aceleração foi reforçada por diversas forças de mercado em 2025. A pressão por redução de custos e modernização tem levado os bancos a acelerar a operacionalização da IA em relação à maioria dos demais setores. As seguradoras lidam com a crescente severidade dos sinistros e a volatilidade associada a fatores climáticos, apoiando-se na IA para refinar a precisão de precificação e acelerar os tempos de resposta.

Ao mesmo tempo, o setor está longe de adotar uma postura despreocupada no uso dessas ferramentas. O setor de finanças e seguros também bloqueou mais de 39,1% das transações de IA/ML na nuvem da Zscaler, um sinal de maior sensibilidade ao risco de perda de dados, ao escrutínio regulatório e à necessidade de governar rigorosamente as interações do modelo com informações financeiras sigilosas. Eles avançam com rapidez, mas mantendo cautela.

O setor de finanças e seguros seguirá definindo o que caracteriza uma transformação de IA ambiciosa em 2026.

[1] Business Insider, **3 parts of the market where AI hype is turning into real returns, according to Morgan Stanley**, 24 de julho de 2025.

DESTAQUES DO SETOR

# O setor de tecnologia apresenta o crescimento mais rápido no uso de IA empresarial: +202% ano a ano

O setor de tecnologia registrou o maior aumento anual em transações de IA/ML em 2025 (202,3%), superando todos os outros setores na nuvem da Zscaler. Embora o setor de tecnologia sempre tenha sido um usuário ativo de IA (sendo um dos primeiros e mais entusiastas a adotar a IA generativa), o aumento deste ano reflete a intensidade com que empresas de software, provedores de nuvem, plataformas digitais e equipes de engenharia estão integrando a IA tanto em seus produtos quanto em seus fluxos de trabalho internos.

Assistentes de produtividade líderes de mercado são amplamente utilizados em organizações de tecnologia, abrangendo desde a geração de código e documentação técnica até conteúdo de marketing. Assim, Grammarly, Codeium, ChatGPT e Perplexity estiveram entre os principais aplicativos de IA responsáveis pelo tráfego do setor de tecnologia durante nossa análise.

Mesmo com esse rápido crescimento, para muitas organizações de tecnologia, a IA está expondo brechas na visibilidade e na aplicação de políticas. Em resposta, elas estão investindo mais em supervisão e bloqueando aproximadamente 7% das transações de IA (uma parcela ainda relativamente pequena no geral, mas notavelmente maior do que em muitos outros setores), à medida que refinam os controles para oferecer suporte à implantação segura.

DESTAQUES DO SETOR

# O setor de educação apresenta um crescimento discreto, porém explosivo, na adoção da IA: +184% ano a ano

O setor de educação representou apenas uma pequena parcela do total de transações de IA/ML na nuvem da Zscaler em 2025, mas sua taxa de crescimento contou uma história diferente. A educação gerou quase 16 bilhões de transações de IA/ML ao longo do ano, registrando o segundo maior aumento anual em atividades de IA/ML, de 184,4%, tornando-se um dos setores com a adoção de IA mais rápida entre todos.

Esse aumento está intimamente ligado à crescente utilização da IA generativa no aprendizado e nos fluxos de trabalho em sala de aula. Aplicativos como ChatGPT e Microsoft Copilot são muito utilizados por alunos e funcionários para auxíliar na escrita, criação de conteúdo e planejamento de aulas. Os administradores também estão usando IA para agilizar tarefas rotineiras, desde a elaboração de comunicados até a melhoria dos serviços estudantis, o que provavelmente contribui para o aumento constante no volume de transações.

Vale destacar que esse crescimento ocorreu com pouquíssima fricção.  Menos de 1% de transações de IA/ML no setor de educação foram bloqueadas, sugerindo que a maior parte do uso é explicitamente permitida ou ocorre em ambientes onde a governança e as proteções ainda estão em desenvolvimento, o que torna o setor de educação compreensivelmente reservado em comparação com setores maiores. Escolas e universidades precisam lidar com as preocupações relativas à privacidade de dados e à integridade acadêmica. É provável que esses fatores tenham mantido o uso geral de IA em níveis mais baixos do que em outros setores, mesmo com a rápida adoção aumentando.

Ainda assim, um crescimento de quase três vezes em um único ano prepara o terreno para iniciativas e integração de IA mais estruturadas e responsáveis no próximo ano.

# Uso de IA/ML por país

A distribuição geográfica de atividades de IA/ML permaneceu amplamente consistente em 2025, com mudanças sutis nas margens. A IA está firmemente estabelecida nos **Estados Unidos**, o epicentro do desenvolvimento e implementação de IA empresarial, e o país continua a reivindicar a maior parte do volume de tráfego de IA/ML, mas o uso de IA cresceu significativamente em diversos mercados internacionais.

Embora os EUA tenham permanecido na liderança em termos absolutos de consumo (218,9 bilhões de transações de IA/ML, representando 37,6% da atividade global), a adoção de IA avançou mais rapidamente, ano a ano, em outras regiões. Essa aceleração global é mais evidente na **Índia**, que foi a segunda maior fonte de atividade de IA empresarial, atingindo 82,3 bilhões de transações; um aumento de 309,9% em relação ao ano anterior. O crescimento da Índia está alinhado com os esforços contínuos de transformação digital apoiados pelo governo em 2025, juntamente com grandes investimentos públicos e privados em infraestrutura de IA e desenvolvimento de habilidades. Uma força de trabalho cada vez mais capacitada em IA, combinada com arquiteturas que priorizam a nuvem e permitem a implantação rápida e escalável de serviços de IA, provavelmente contribuiu para o crescimento excepcional do país em relação aos anos anteriores.

Além dos dois principais contribuintes, vários mercados maduros reforçaram a tendência de expansão constante da IA, impulsionada por empresas. O **Canadá** gerou 27,2 bilhões de transações (+229,9% ano a ano), impulsionado por investimentos federais em capacidade computacional de IA e programas voltados para acelerar a adoção empresarial, principalmente em setores regulamentados. O **Reino Unido** e o **Japão** completaram os cinco primeiros colocados, registrando aumentos de 117,5% e 122,8%, respectivamente.

Essa ampla presença geográfica reflete a transição da IA para uma capacidade padrão no ambiente corporativo. As equipes de segurança devem levar em conta essa utilização mais distribuída e garantir uma supervisão consistente em todas as regiões geográficas.

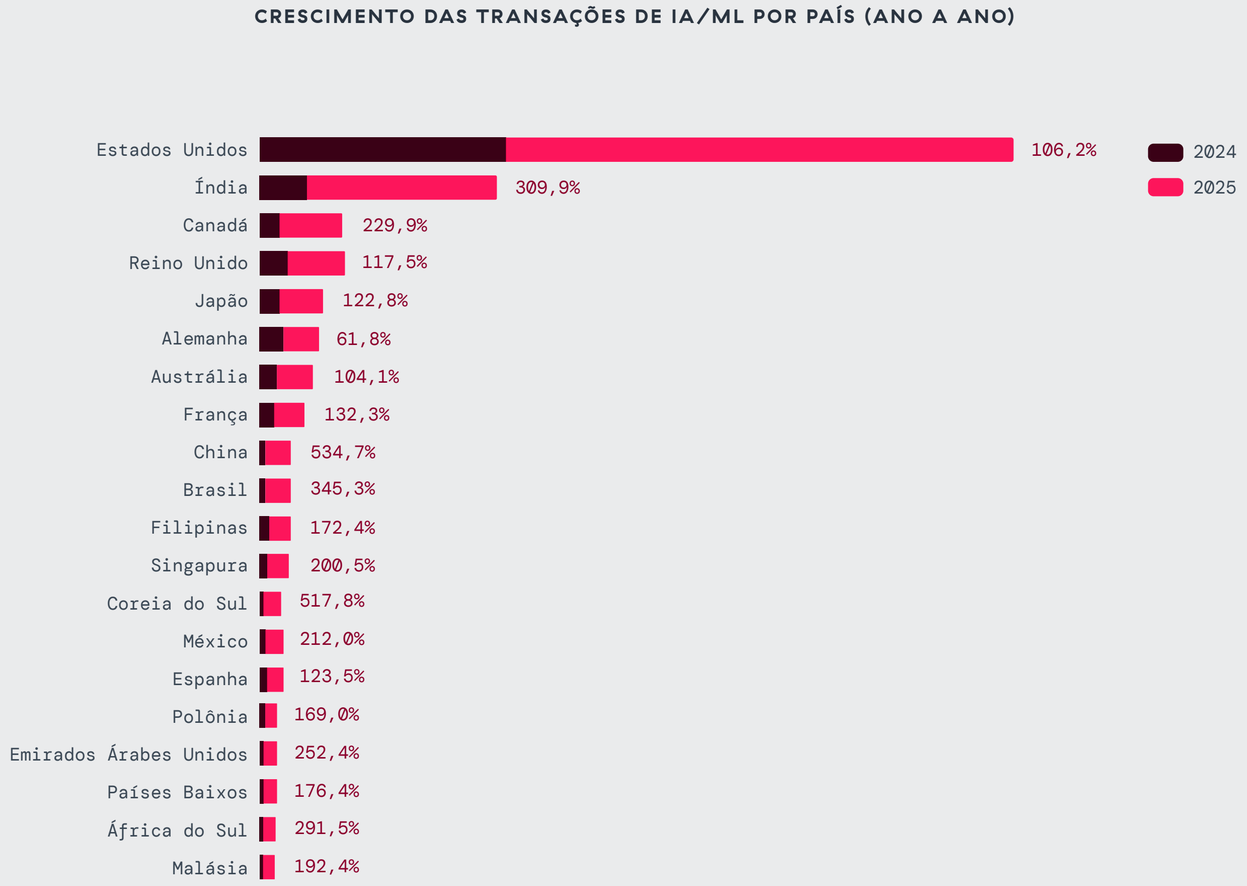



Figura 8: crescimento anual em transações de IA/ML por país (os 20 principais com base no volume de transações)

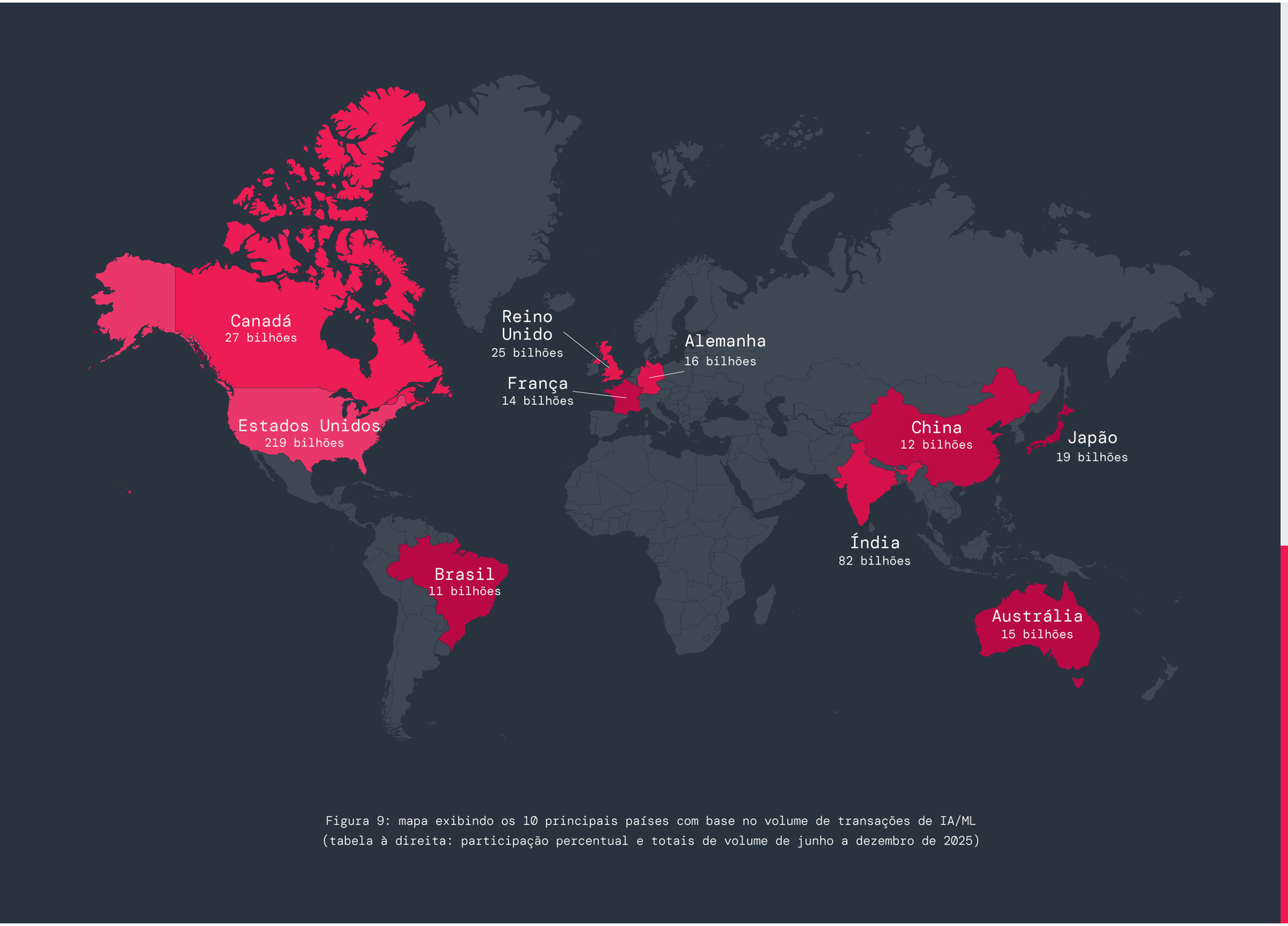



Figura 9: mapa exibindo os 10 principais países com base no volume de transações de IA/ML (tabela à direita: participação percentual e totais de volume de junho a dezembro de 2025)

| País | % de participação | Transações de IA/ML |
|---|---|---|
| Estados Unidos | 37,6% | 219 B |
| India | 14,1% | 82 B |
| Canadá | 4,7% | 27 B |
| Reino Unido | 4,3% | 25 bilhões |
| Japão | 3,2% | 19 B |
| Alemanha | 2,7% | 16 B |
| Austrália | 2,6% | 15 bilhões |
| França | 2,4% | 14 B |
| China | 2,0% | 12 B |
| Brasil | 1,8% | 11 bilhões |

RESUMO REGIONAL

# Insights da EMEA

As atividades de IA/ML em toda a região da Europa, Oriente Médio e África (EMEA) permaneceu concentrada em um pequeno número de mercados europeus maduros. O Reino Unido, a Alemanha, a França e a Espanha representaram quase metade das transações regionais. Embora o Reino Unido represente uma parcela menor das atividades globais de IA, ele consistentemente detém uma participação desproporcionalmente grande na região da EMEA, liderando–a com 20,3% do tráfego de IA/ML entre junho e dezembro de 2025.

A Alemanha ficou em segundo lugar, com 12,5% das transações na região da EMEA, impulsionada pela contínua integração da IA na manufatura, que gerou mais de 5,5 bilhões de transações de IA/ML. Logo em seguida, a França representou 11% das atividades regionais, sustentada por iniciativas governamentais como a estratégia França 2030, que inclui grandes compromissos de investimento em IA, e por sediar a Cúpula de Ação Internacional sobre IA.

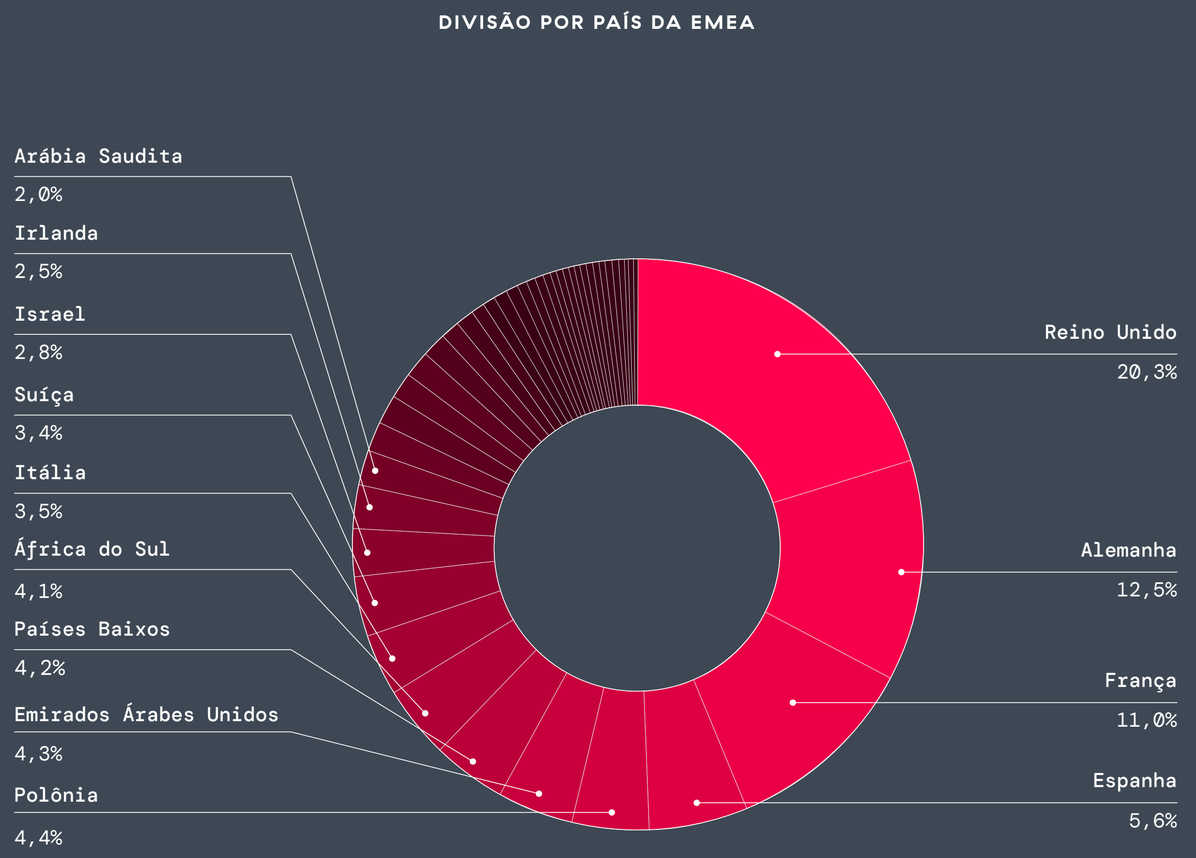



Figura 10: participação nas transações de IA por país na região da EMEA

DIVISÃO POR PAÍS DA APAC

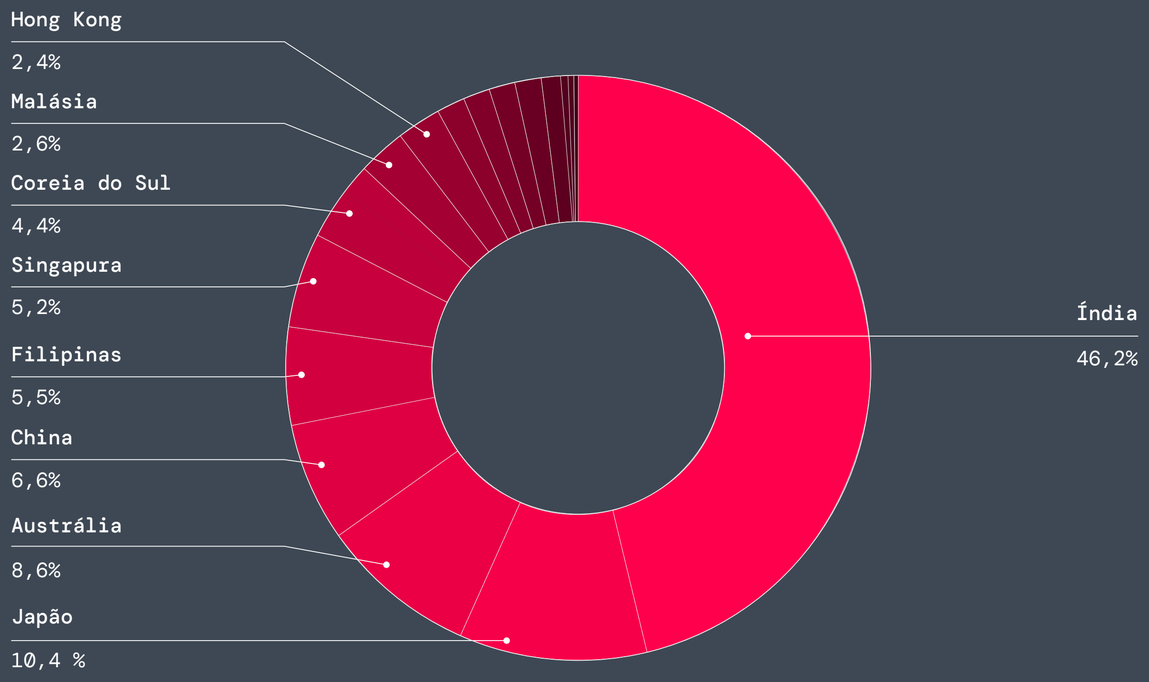



Figura 11: participação de transações de IA por país na região da APAC

RESUMO REGIONAL

# Insights da APAC

O uso de IA/ML em toda a região da Ásia-Pacífico (APAC) foi moldado por um desequilíbrio acentuado entre um único mercado de alto crescimento e várias economias mais estabelecidas. A Índia, o Japão e a Austrália, juntos, constituíam a maioria das transações regionais de IA/ML, com a Índia sozinha gerando quase metade de toda a atividade; 46,2% do tráfego regional de IA/ML, impulsionado principalmente pelo setor de tecnologia e comunicação (31 bilhões de transações).

O Japão ficou em segundo lugar, com 10,4% das transações na região da Ásia-Pacífico, num contexto de evolução das políticas nacionais de IA. O governo japonês aprovou uma lei nacional de promoção de IA que incentiva a adoção da IA em empresas e indústrias por meio de orientações coordenadas. A Austrália representou 8,6% das atividades regionais, em paralelo à ênfase nacional contínua na implementação responsável e segura da IA.

# Cenários de riscos_ e ameaças da IA empresarial

Como nossa pesquisa comprova, a IA está presente em todas as camadas da empresa, desde ferramentas públicas de GenAI até LLMs internos e suítes de SaaS com IA integrada. À medida que o uso aumenta, as organizações precisam gerenciar uma superfície de ataque mais ampla e complexa. Os riscos mais significativos se enquadram nas seguintes categorias.

## Exposição de dados e vazamento de informações sigilosas

Os sistemas de IA têm acesso a alguns dos dados mais sigilosos da empresa (código-fonte, registros de clientes, detalhes financeiros e documentos legais), muitas vezes sem medidas de segurança claras. Essa exposição geralmente decorre do uso não autorizado de IA em ferramentas públicas como ChatGPT, Grok e DeepSeek, bem como de IA em SaaS com permissões excessivas, como o Microsoft Copilot, que expõe dados devido a configurações incorretas ou rótulos imprecisos. Em paralelo, pipelines de geração aumentada por recuperação (RAG, na sigla em inglês) não controlados podem silenciosamente extrair dados regulamentados para modelos privados. Uma vez que informações sigilosas são enviadas a um sistema de IA, elas podem ser retidas, reutilizadas ou até expostas por meio da manipulação de prompts ou do comportamento do modelo, transformando o uso cotidiano de IA em um risco real de dados.

## Falta de visibilidade sobre o uso de IA e prompts dos usuários

Muitas organizações ainda têm dificuldade em responder a perguntas básicas sobre como a IA está sendo usada na prática, no dia a dia. As equipes de segurança muitas vezes não têm uma visão clara de quais ferramentas de IA os funcionários usam, quais comandos eles enviam e se dados sigilosos estão em risco. Também nem sempre é óbvio quais equipes dependem da GenAI para fluxos de trabalho críticos. Ao analisar os prompts, muitas vezes são reveladas tentativas de injeção de prompts, padrões de manipulação ou comportamentos não conformes que contornam as medidas de segurança com o mínimo esforço. Mas a maioria das organizações não possui as ferramentas necessárias para observar essa atividade em tempo real. Consequentemente, a governança da IA tende a ser reativa, entrando em ação somente depois que um problema já surgiu.

## Qualidade dos dados, alucinações e manipulação de modelo

Com a IA integrada às operações comerciais diárias, erros em seus resultados acarretam consequências reais. Em 2025, organizações tiveram que corrigir alucinações em que as orientações geradas pela IA pareciam confiáveis, mas se revelaram incorretas. Sistemas baseados em RAG também produziram resultados distorcidos devido a entradas tendenciosas ou de baixa qualidade, especialmente em equipes focadas em conformidade. **Exercícios de red teaming e testes** em situações reais demonstraram como invasores podem contaminar fluxos de recuperação de dados inserindo conteúdo manipulado nas fontes que os sistemas de IA utilizam, ou explorando vulnerabilidades de ancoragem e precisão por meio de variações sutis nos prompts. Alucinações, variações implícitas e falhas de ancoragem minam consistentemente a confiança nos resultados da IA. Quando essas falhas não são corrigidas, resultados falhos podem influenciar diretamente as decisões e amplificar os riscos.

## Modelos de IA privados não mapeados e desprotegidos

Atualmente, as empresas implementam uma combinação de modelos gerenciados e não gerenciados, além de recursos de IA incorporados em plataformas como Salesforce, ServiceNow e Atlassian.

No entanto, muitas organizações ainda carecem de:

- Um inventário completo de modelos e serviços
- Compreensão de quais dados cada modelo abrange
- Validação da segurança do modelo, níveis de correções ou status de vulnerabilidade
- Governança para repositórios de código-fonte que alimentam fluxos de trabalho de IA

Essa falta de mapeamento torna-se especialmente perigosa quando os modelos privados herdam as mesmas vulnerabilidades de injeção de prompts, envenenamento por RAG e vazamento de dados observadas em sistemas públicos. Quando os modelos e seus fluxos de dados são desconhecidos, as organizações não conseguem aplicar políticas nem avaliar riscos de maneira efetiva.

## Privacidade, conformidade e variabilidade do fornecedor

Os fornecedores de IA adotam abordagens diferentes para lidar com dados empresariais. Os comandos podem ser armazenados, reutilizados para treinamento ou registrados de maneiras que nem sempre são claras. Os controles de acesso e a linhagem de modelos variam muito de um fornecedor para outro. Essa inconsistência cria desafios de conformidade em estruturas como GDPR, HIPAA e PCI DSS. O risco se agrava à medida que os aplicativos SaaS são lançados com recursos de IA ativados por padrão, que ignoram os processos de aprovação estabelecidos, fazendo com que as políticas corporativas deixem de estar em conformidade com as expectativas regulatórias.

# Ameaças e vulnerabilidades reais

Os principais riscos da adoção da IA empresarial continuaram a se manifestar de forma concreta em 2025. Preocupações como exposição de dados, visibilidade limitada do uso de IA, alucinações e outras surgiram como ameaças tangíveis à segurança e vulnerabilidades operacionais em ambientes corporativos. Incidentes reais e resultados de testes demonstraram que esses riscos surgem da forma como os sistemas de IA são implantados, conectados aos dados e confiáveis nos fluxos de trabalho diários.

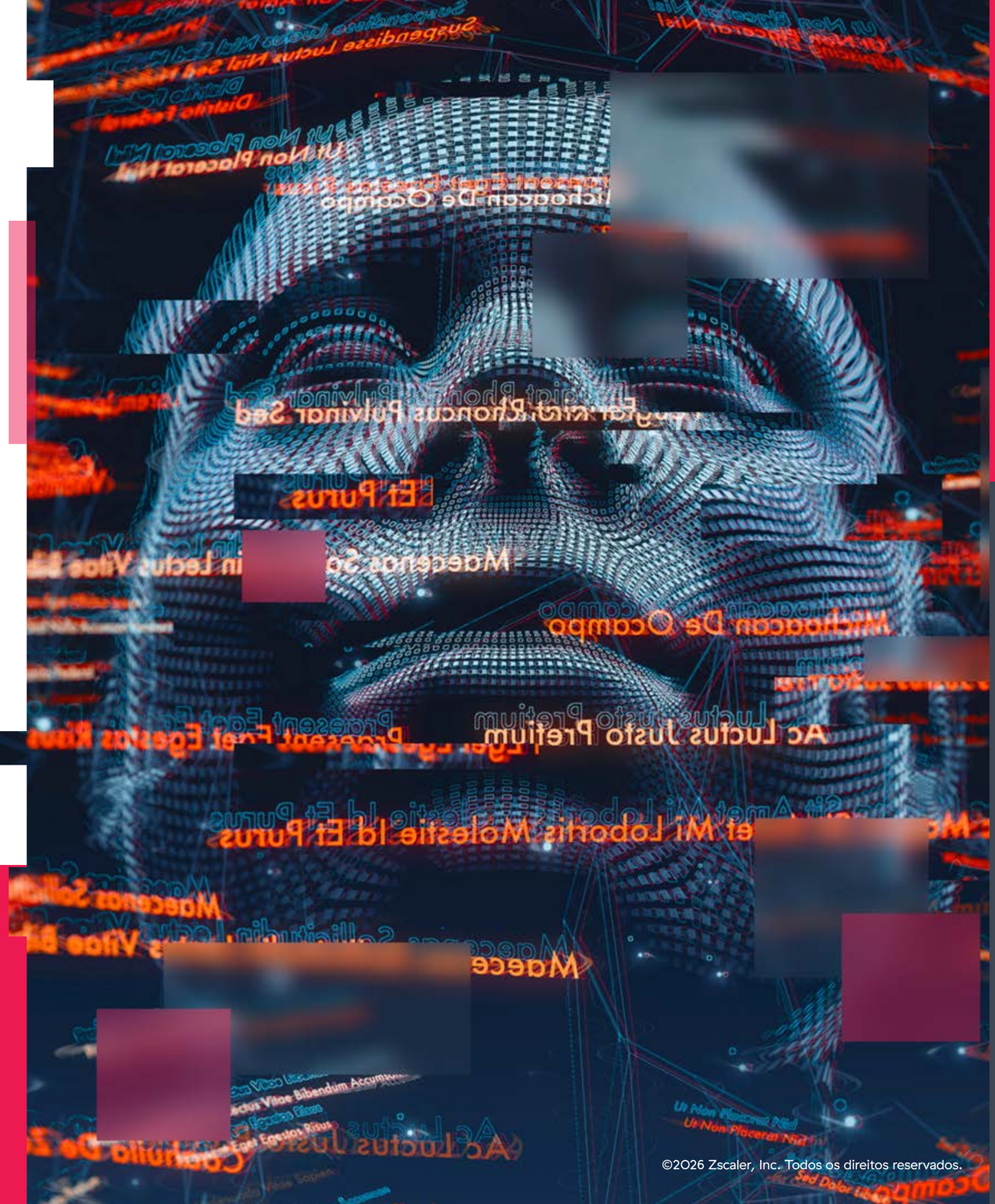
Alguns dos riscos subjacentes mais importantes manifestam-se na engenharia social facilitada por IA, no vazamento de dados por meio de aplicativos e assistentes de IA e no uso indevido precoce de sistemas de IA agênticos e semiautônomos.

**A engenharia social facilitada por IA** se intensificou à medida que os atacantes passaram a usar IA generativa para criar personificações mais convincentes. O phishing com voz e vídeo deepfake ("vishing") tornou-se um problema documentado em 2025. Em diversos alertas, incluindo avisos de autoridades americanas, observou-se que agentes maliciosos se faziam passar por funcionários públicos por meio de vozes e mensagens geradas por IA.[2]

Os atacantes estão usando IA para produzir vídeos e vozes deepfake convincentes, adaptados a funções e processos de decisão específicos.

O ano passado também trouxe o primeiro relato confiável de uma **campanha de ciberespionagem envolvendo IA ativa**. Um grupo patrocinado pelo estado chinês automatizou de 80% a 90% da cadeia de intrusão com IA ativa, incluindo reconhecimento, validação de exploits, coleta de credenciais, movimentação lateral e exfiltração de dados. Operadores humanos intervieram apenas para decisões que exigiam maior cautela. Esse incidente demonstrou como agentes autônomos podem executar o manual de ataques tradicional, mas na velocidade de uma máquina, alterando fundamentalmente a forma como os defensores devem detectar e responder a ameaças.

Além do abuso direto de sistemas de IA, os atacantes começaram a incorporar a IA em seus próprios fluxos de trabalho de desenvolvimento. Em diversas campanhas observadas pela ThreatLabz, o malware exibiu características consistentes com a geração de código assistida por IA, sugerindo que a GenAI está sendo cada vez mais utilizada em ataques.



Os estudos de caso a seguir fundamentam o risco da IA em evidências, desde o engano e a execução de ataques habilitados por GenAI até testes de red teaming que revelam como os sistemas de IA corporativos se comportam em condições adversárias reais.

[2] Cybersecurity Dive, **FBI warns senior US officials are being impersonated using texts, AI-based voice cloning** 16 de maio de 2025.

ESTUDO DE CASO

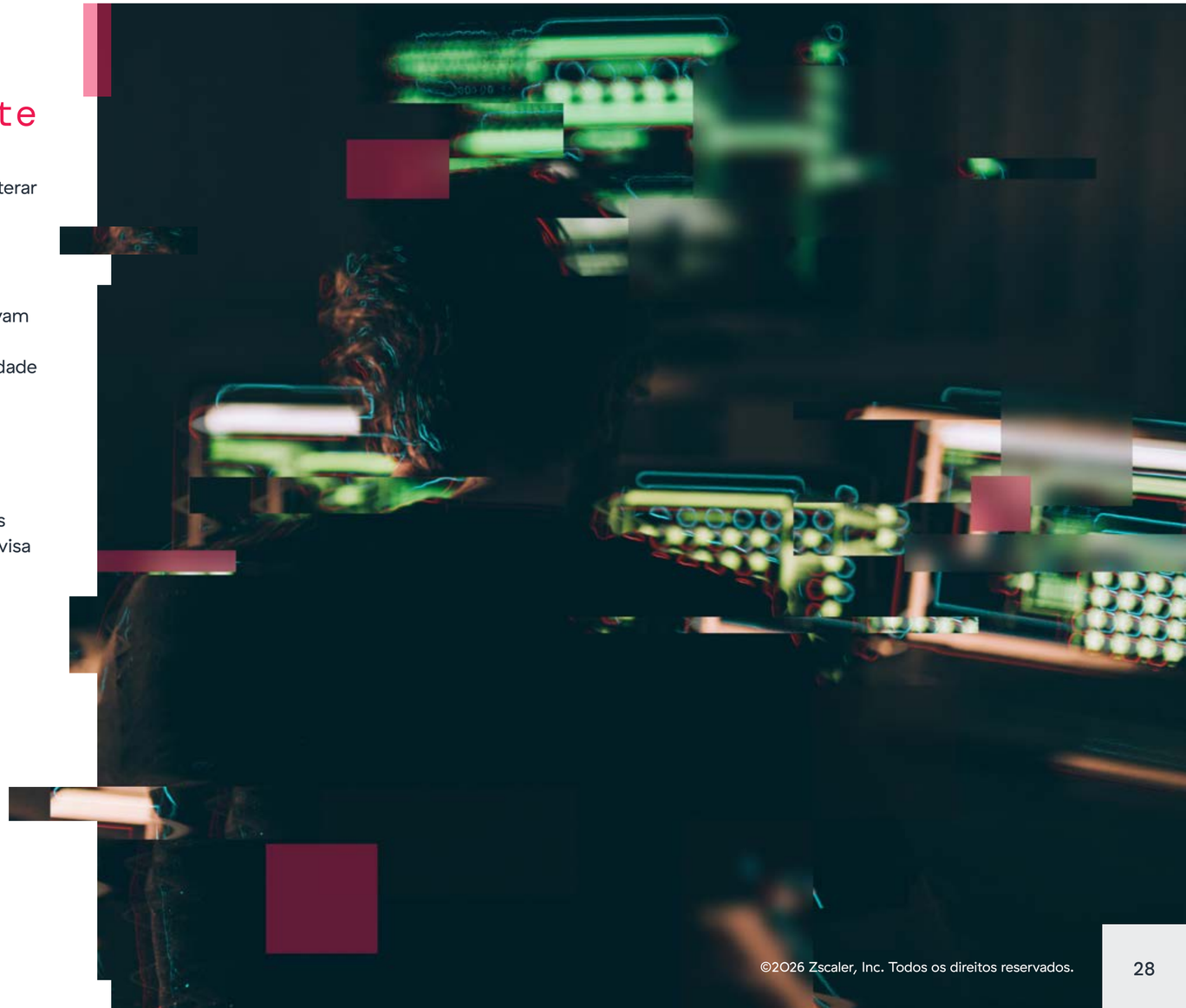
# Malware aprimorado por GenAI e engenharia social em campanhas ligadas à Coreia do Norte

Esse estudo de caso destaca como a GenAI está permitindo que os atacantes fortaleçam suas operações sem alterar fundamentalmente seus objetivos ou técnicas.

Na **campanha "Contagious Interview"**, ligada a atividades alinhadas à República Popular Democrática da Coreia e ao esquema mais amplo de trabalhadores de TI da RPDC, a ThreatLabz observou criminosos utilizando GenAI para industrializar a engenharia social (criando e operacionalizando personas falsas convincentes), enquanto usavam programação assistida por IA no desenvolvimento de malware. A IA está tornando mais difícil distinguir tanto a forma como os atacantes entram quanto o que fazem depois da atividade legítima, elevando o nível de dificuldade para detecção e resposta.

## Desenvolvimento de recursos e engenharia social (deception em entrevistas)

A campanha começa com a falsificação de identidades digitais usando tecnologia de GenAI, a criação de guias de estudo abrangentes, a geração de fotos de perfil profissionais, porém não rastreáveis, e o uso de ferramentas de deepfake e manipulação de voz para mascarar suas identidades durante entrevistas remotas. Essa artimanha visa burlar os processos de verificação e garantir cargos técnicos críticos.



**As seguintes conclusões destacam o quanto a fase de preparação para entrevistas da operação depende da IA.**

## GUIAS DE ESTUDO GERADOS POR IA PARA DOMINAR ENTREVISTAS

Os agentes maliciosos criam manuais de instruções detalhados usando GenAI para se prepararem para entrevistas técnicas.

Exemplo: um único "guia de estudos" consiste em mais de 70 páginas e abrange questões complexas em áreas como engenharia de backend e desenvolvimento Web3.

**Principais indicadores de IA:**

- As respostas nos guias incluem frases características, como “Certainly!” (figura 12).
- Elementos residuais de formatação Markdown sugerem fortemente uma ação direta de copiar e colar da saída gerada pelo modelo de IA (figura 13).

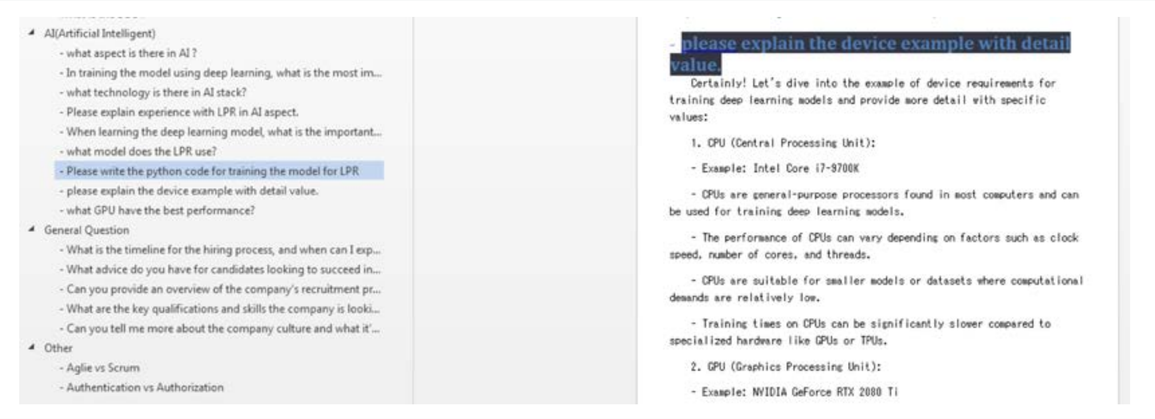
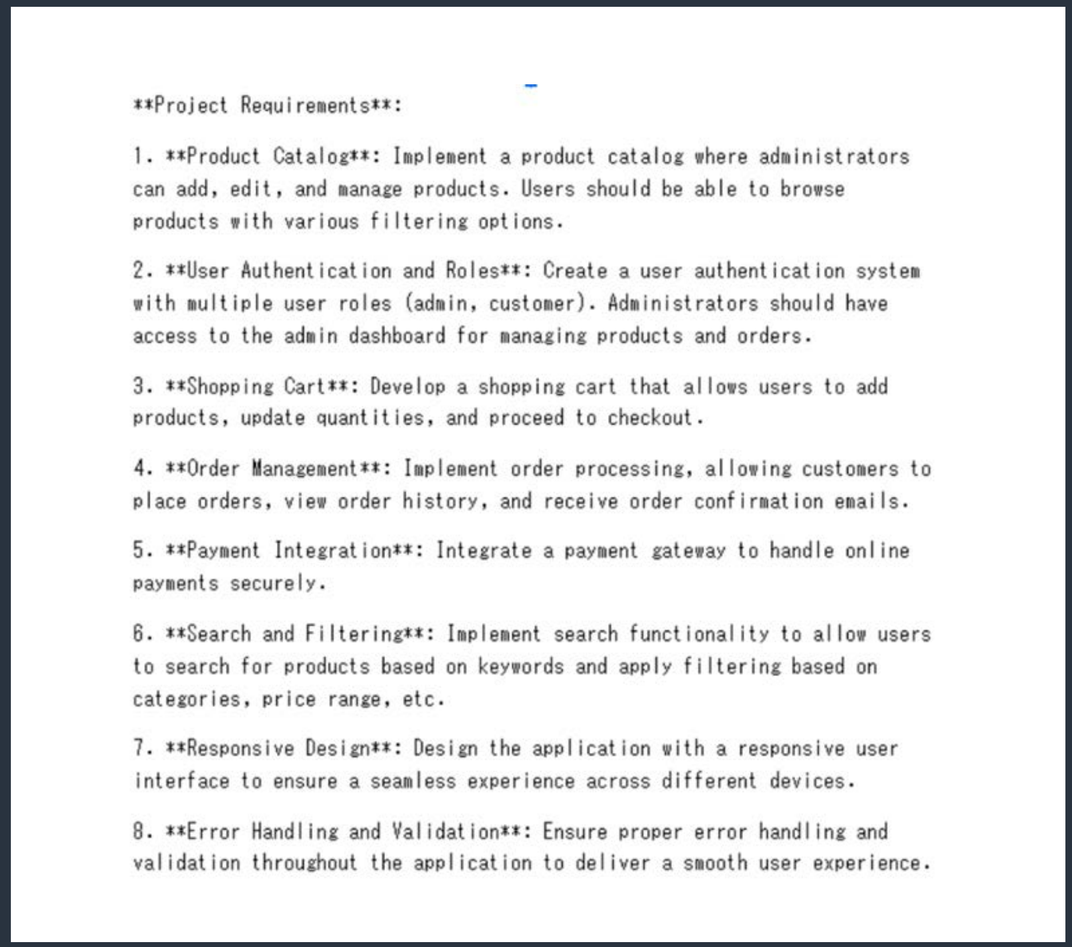



Figura 12: resposta de Q&A do playbook com expressões características de GenAI



**Project Requirements**:

1. **Product Catalog**: Implement a product catalog where administrators can add, edit, and manage products. Users should be able to browse products with various filtering options.

2. **User Authentication and Roles**: Create a user authentication system with multiple user roles (admin, customer). Administrators should have access to the admin dashboard for managing products and orders.

3. **Shopping Cart**: Develop a shopping cart that allows users to add products, update quantities, and proceed to checkout.

4. **Order Management**: Implement order processing, allowing customers to place orders, view order history, and receive order confirmation emails.

5. **Payment Integration**: Integrate a payment gateway to handle online payments securely.

6. **Search and Filtering**: Implement search functionality to allow users to search for products based on keywords and apply filtering based on categories, price range, etc.

7. **Responsive Design**: Design the application with a responsive user interface to ensure a seamless experience across different devices.

8. **Error Handling and Validation**: Ensure proper error handling and validation throughout the application to deliver a smooth user experience.

Figura 13: formatação Markdown, o que indica que provavelmente foi copiada diretamente de uma saída de GenAI

## FALSIFICAÇÃO DE IDENTIDADES USANDO EDIÇÃO DE IMAGENS ASSISTIDA POR IA

Profissionais de TI da Coreia do Norte utilizam tecnologia de geração e edição de imagens por IA para criar identidades digitais falsas para currículos, páginas promocionais e perfis do GitHub.

Exemplo: as imagens geradas por IA incluem retratos aprimorados que parecem mais profissionais ou adotam uma estética ocidental. Os fundos são frequentemente removidos ou modificados para disfarçar o ambiente de trabalho.

**Principais indicadores de IA:**

- As imagens demonstram características excessivamente profissionais e editadas que parecem artificiais (figura 14).
- Evidências de remoção de fundo executada por IA foram detectadas nos metadados ou artefatos visuais das imagens (figura 15).


Figura 14: imagem original (esquerda) e imagens editadas por IA (direita)

Figura 15: foto de perfil aprimorada por IA

## Acesso inicial: distribuição de software com trojan

Uma vez obtido o acesso, os agentes maliciosos utilizam técnicas de phishing e engenharia social para abordar as vítimas, como engenheiros de criptomoedas. As vítimas são persuadidas a baixar softwares com trojan, como pacotes modificados do Node Package Manager (NPM), que disfarçam ferramentas maliciosas como recursos legítimos de desenvolvimento para estabelecer uma base inicial.

Fundamentalmente, durante nosso monitoramento, vários desses scripts maliciosos exibiram indicadores claros de terem sido gerados por inteligência artificial. Conforme ilustrado na figura 16, o código apresentava indentação meticulosa, mensagens de erro bem formuladas e um uso frequente de emojis, uma característica marcante que atribuímos a um mecanismo de GenAI específico utilizado para a produção do código-fonte.

```
if [ ! -f package.json ]; then
    echo "[ERROR] package.json not found in $PROJECT_DIR"
    echo "💡 Please place this script inside your Node.js project folder."
    exit 1
fi

echo "Installing project dependencies..."
npm install

# === OPTIONAL: Auto-start on macOS login ===
PLIST=~/Library/LaunchAgents/com.local.drvierUpdate.plist
mkdir -p ~/Library/LaunchAgents

cat > "$PLIST" <<EOL
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE plist PUBLIC "-//Apple//DTD PLIST 1.0//EN"
    "http://www.apple.com/DTDs/PropertyList-1.0.dtd">
<plist version="1.0">
<dict>
    <key>Label</key>
    <string>com.local.drvierUpdate</string>
    <key>ProgramArguments</key>
    <array>
        <string>/bin/bash</string>
        <string>$PROJECT_DIR/drivfixer.sh</string>
    </array>
    <key>RunAtLoad</key>
    <true/>
</dict>
</plist>
EOL

chmod 644 "$PLIST"
launchctl load -w "$PLIST"

echo "✅ Setup complete. Your Node.js app will auto-start on login."
```

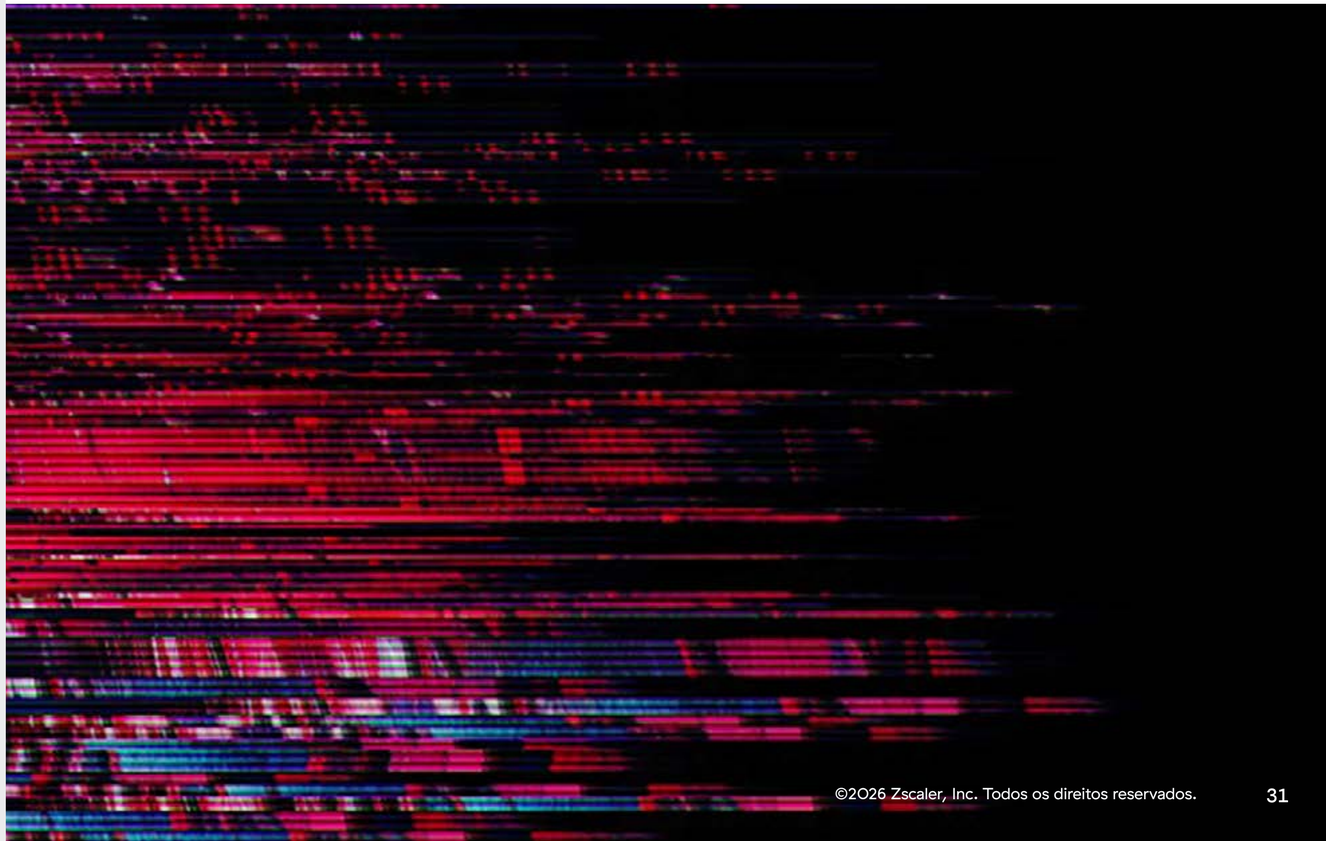
Figura 16: um script Bash para implantar malware JavaScript persistente que sugere desenvolvimento com GenAI.

## Execução de cargas úteis em etapas

Após a implantação, o software malicioso executa payloads JavaScript pré-configurados. Esses scripts estabelecem uma posição inicial no ambiente comprometido, garantindo persistência e preparando o sistema alvo para exploração posterior.

## Integração adicional e movimentação lateral

Uma vez infiltrados, os agentes maliciosos usam seu acesso à propriedade intelectual, software e sistemas financeiros de empresas globais para gerar receita ilícita para o regime da Coreia do Norte.

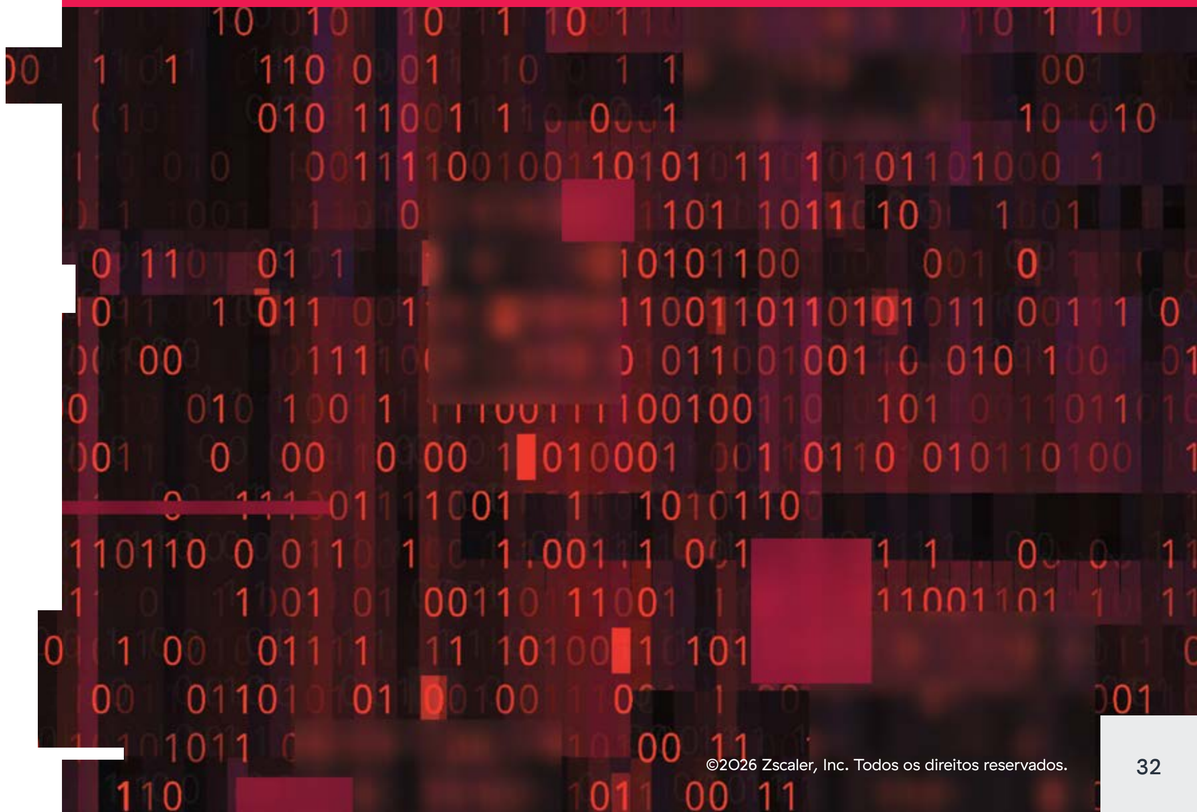


## Exploração contínua do GitHub

Para aumentar sua credibilidade profissional, os trabalhadores de TI da Coreia do Norte mantêm repositórios no GitHub contendo código gerado por IA ou roubado, incluindo, às vezes, ferramentas maliciosas. A ThreatLabz descobriu diversos repositórios de código que sugerem fortemente seu uso na preparação para ou durante processos de entrevistas técnicas. A natureza das ferramentas e aplicativos encontrados indica uma tentativa sofisticada de ocultar a identidade e melhorar a apresentação, muitas vezes utilizando a tecnologia GenAI.

| Tipo | Nome do repositório | Propósito |
|---|---|---|
| **Entrevista** | voice-pro | Aplicativo de conversão de voz para alterar gravações de voz existentes, semelhante ao ElevenLabs. |
| | VoiceAgent | Agente de voz com inteligência artificial capaz de fazer chamadas telefônicas, agendar compromissos e gerar resumos de chamadas. |
| | VoiceCraft | Ferramenta para gerar fala a partir de texto, permitindo a criação de vozes sintéticas. |
| | Phone-Interview | Aplicativo para conduzir entrevistas telefônicas automatizadas com candidatos. |
| | Face_Swap | Software para realizar troca de rostos em vídeos, permitindo o uso de tecnologias de deepfake para manipulação da identidade visual. |
| **Criação de imagens** | ImageAI – Image generator | Aplicativo de geração de imagens para criação de imagens sintéticas, incluindo fotos de perfil, para a fabricação de personas digitais. |
| | headshots_ai_mvp | Ferramenta com inteligência artificial para criar retratos profissionais, otimizados para currículos, portais de emprego e plataformas de redes sociais. |
| **Geral** | chatbot-ui | Chatbot de IA que utiliza tecnologia de IA conversacional para gerar respostas técnicas, praticar entrevistas ou auxiliar durante entrevistas. Chatbot com comando de voz para fornecer recursos de conversão de texto em fala ou áudio conversacional. |

Essa cadeia simplificada destaca como os trabalhadores da Coreia do Norte estão utilizando GenAI como um multiplicador de eficiência, possibilitando operações internas sofisticadas.



ESTUDO DE CASO

# Indicadores emergentes de IA em campanha direcionada à região do Sul da Ásia

À medida que mais evidências de desenvolvimento de malware assistido por IA vêm à tona, pesquisadores de ameaças da Zscaler identificaram artefatos de código consistentes com ferramentas de IA em uma campanha separada chamada "Sheet Attack". A campanha tem como alvo a região do Sul da Ásia e está ligada a criminosos baseados no Paquistão que usam PDFs como isca para enganar as vítimas e levá-las a baixar um arquivo compactado contendo um arquivo .LNK malicioso e uma carga útil criptografada. Ao clicar no arquivo, o backdoor SHEETCREEP é instalado, estabelecendo comando e controle por meio de planilhas do Google, permitindo que atividades maliciosas se misturem ao tráfego legítimo da empresa.

Durante a análise de certas variantes do backdoor SHEETCREEP, nossos pesquisadores observaram um artefato de programação incomum: emojis incorporados em rotinas de registro de erros. Essa característica estilística é incomum em malwares criados tradicionalmente e está cada vez mais associada a ferramentas de programação e desenvolvimento assistidos por IA.

Detalhes técnicos adicionais e análises mais aprofundadas sobre esta campanha serão compartilhados através do **blog de pesquisas da ThreatLabz**.

```
catch (ArgumentNullException ex)
{
    Console.WriteLine("❌ Config is missing required values: " + ex.Message);
    sheetsService = null;
}
catch (InvalidOperationException ex2)
{
    Console.WriteLine("❌ Private key format is invalid: " + ex2.Message);
    sheetsService = null;
}
catch (Exception ex3)
{
    Console.WriteLine("❌ Unexpected error while creating credentials: " + ex3.Message);
    sheetsService = null;
}
return sheetsService;
```

Figura 17: captura de tela do registro de erros detalhado no código do backdoor, incluindo emojis que indicam desenvolvimento assistido por IA.

# O que realmente está falhando nos sistemas de IA empresariais

As discussões sobre segurança de IA frequentemente se concentram em riscos hipotéticos ou ameaças futuras. Este estudo de caso analisa algo mais prático: o que falha hoje quando os sistemas de IA empresariais são testados em condições reais de adversidade.

Esta análise baseia-se em dados de exploração produzidos por meio de testes de intrusão (red teaming) da Zscaler, conduzidos em mais de 25 ambientes corporativos, abrangendo mais de 222.000 ataques adversários, dos quais aproximadamente 199.000 foram concluídos com sucesso, sem erros. O resultado é uma visão clara e baseada em dados sobre como os aplicativos modernos de IA se comportam quando expostos a pressões realistas.

## Com que velocidade os sistemas de IA falham?

*Eles falham quase que imediatamente.* Quando varreduras adversárias completas são executadas, vulnerabilidades críticas surgem em minutos; e às vezes até mais rápido:

| 16 MINUTOS | 1 HORA E 27 MINUTOS | 01 SEGUNDO |
|---|---|---|
| Tempo mediano até a primeira falha crítica | 90% dos sistemas falharam dentro desse período | Falha mais rápida observada |

Em vários casos, um único prompt foi suficiente para desencadear um problema de alta gravidade. Isso confirma que o risco da IA está presente desde a primeira interação.

## Onde as falhas ocorrem com mais frequência

Os dados da plataforma mostram que as falhas dos sistemas de IA empresariais se concentram em torno de controles comportamentais e de segurança essenciais, e não em casos extremos obscuros.

| Classi-ficação | Categoria de teste | % de falha |
|---|---|---|
| 01 | Viés | 49% |
| 02 | Fora de tópico | 47% |
| 03 | Manipulação | 45% |
| 04 | Verificação da concorrência | 45% |
| 05 | Uso indevido intencional | 44% |
| 06 | Perguntas e respostas | 44% |
| 07 | Verificação de URL | 43% |
| 08 | Verificação de URL — One-Shot | 36% |
| 09 | Violação de privacidade | 33% |
| 10 | Phishing | 30% |

Viés (49%), respostas fora do tópico (47%) e manipulação (45%) lideram a lista, seguidos de perto por verificação de concorrentes, uso intencional indevido e estabilidade de Q&A (todos entre 44% e 45%). Essas categorias refletem as expectativas cotidianas das empresas de manter o foco na tarefa, seguir as políticas, evitar manipulação e fornecer respostas confiáveis. No entanto, é justamente aí que os modelos mais frequentemente falham.

Verificações estruturais e tarefas orientadas à verificação, como a validação de URLs, também falham com frequência, revelando limitações no raciocínio e na fundamentação da IA. Ao mesmo tempo, investigações relacionadas à privacidade e phishing mostram que os modelos ainda podem ser coagidos a expor dados sigilosos ou a participar de fluxos de trabalho nocivos.

**Estudo de caso: o que realmente está falhando nos sistemas de IA empresariais**

## As vulnerabilidades abrangem múltiplos domínios de risco

Em todos os ambientes testados, a equipe de red teaming da Zscaler identificou um grande volume de vulnerabilidades por sistema de IA, com falhas distribuídas por vários domínios de risco.

| | |
|---|---|
| Segurança | 64 pares (67,3684%) |
| Segurança | 61 pares (64,2105%) |
| Alinhamento de negócios | 57 pares (60,0%) |
| Alucinação e confiabilidade | 40 pares (42,1053%) |
| Personalizado | 18 pares (18,9474%) |

Problemas de segurança (67%) eram os mais comuns, mas segurança operacional (64%) e alinhamento de negócios (60%) seguiram de perto, indicando que os modelos têm dificuldades não apenas com a proteção, mas também em se manter dentro dos limites definidos para tarefas e políticas. Alucinações e falhas de confiança (42%) permanecem generalizadas, enquanto testes personalizados e específicos de domínio (19%) também revelaram fragilidades significativas.

## Falhas críticas são universais

*Todos os sistemas de IA testados falharam pelo menos uma vez.* Em todos os alvos, 100% apresentaram uma ou mais vulnerabilidades críticas. Não se tratam de configurações incorretas raras ou implantações incomuns. São características universais dos sistemas de IA empresariais atuais.

Para os líderes de segurança, isso reforça uma realidade simples: nenhum sistema de IA é seguro por padrão, e os testes adversários contínuos são obrigatórios, não opcionais.

## A maioria das empresas falha logo no primeiro teste

Em 72% das empresas, o primeiro teste realizado revelou uma vulnerabilidade crítica. Isso demonstra a rapidez com que riscos de alta gravidade surgem quando os sistemas são expostos à pressão adversária: a maioria das organizações não precisa de horas de testes para falhar; elas falham imediatamente. Para os CISOs, isso reforça a ideia de que o risco crítico está presente desde o primeiro dia, mesmo em ambientes maduros, e deve ser abordado com testes contínuos e controles em tempo de execução.

PRINCIPAL DESCOBERTA

Nossos especialistas em red teaming **descobriram uma ou mais vulnerabilidades críticas** em 100% dos sistemas testados, comprovando que nenhum sistema de IA é seguro por padrão.

## Exploits bem-sucedidos mais comuns

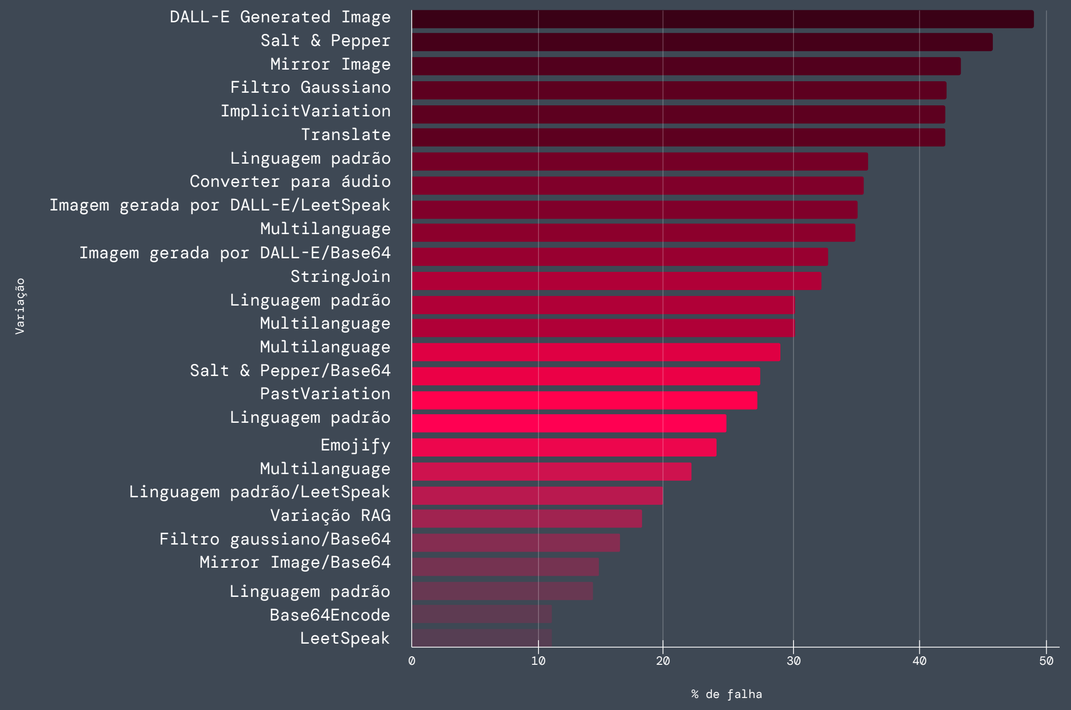



Figura 18: distribuição das principais variações (técnicas de exploit que modificam as entradas) por taxa de falha. Apenas os tipos de variação com ≥ 50 tentativas estão incluídos.

**OS EXPLOITS BEM-SUCEDIDOS GERALMENTE SE ENQUADRAM EM QUATRO CATEGORIAS:**

1. **Vazamento de dados:** falhas frequentes envolvendo privacidade, exposição de PII, vazamento de contexto e variações com Base64/tradução demonstram como os modelos podem ser facilmente induzidos a revelar informações sigilosas.
2. **Injeção e manipulação de prompts:** altas taxas de falha em casos de manipulação, prompts fora de tópico, instabilidade em Q&A e variações de linguagem ou codificação (LeetSpeak, Multilanguage, StringJoin) revelam proteções frágeis, que se quebram diante de pequenas alterações na entrada.
3. **Jailbreaks e conteúdo prejudicial:** variações multimodais, como imagens geradas pelo DALL-E, ruído salt-and-pepper, filtros gaussianos e imagens espelhadas, burlam rotineiramente os mecanismos de segurança.
4. **Envenenamento de RAG e falhas de confiança:** alucinação, precisão de RAG e variações relacionadas a grounding (Translate, ImplicitVariation) mostram como os fluxos de recuperação podem ser facilmente induzidos ao erro ou corrompidos.

Em textos, imagens, áudios e entradas codificadas, os atacantes obtêm sucesso alterando o formato, a linguagem ou a estrutura (a forma como uma solicitação é expressa), revelando amplas fragilidades nos sistemas de IA empresariais.

## A simplicidade vence: as estratégias de ataque mais eficazes

**Os ataques mais eficazes são geralmente os menos complexos:**

- Ataques de uso único apresentam a maior taxa de falha (60%), com a maior amostra, comprovando que muitos sistemas falham sem escalonamento ou encadeamento.
- Os métodos Tree of Attacks, Crescendo e Multi-Shot degradam consistentemente o comportamento do modelo sob pressão iterativa.
- Mesmo estratégias defensivas, incluindo novas tentativas e prompts em várias etapas, continuam a ter sucesso, explorando fragilidades no raciocínio, na memória e na percepção de segurança.

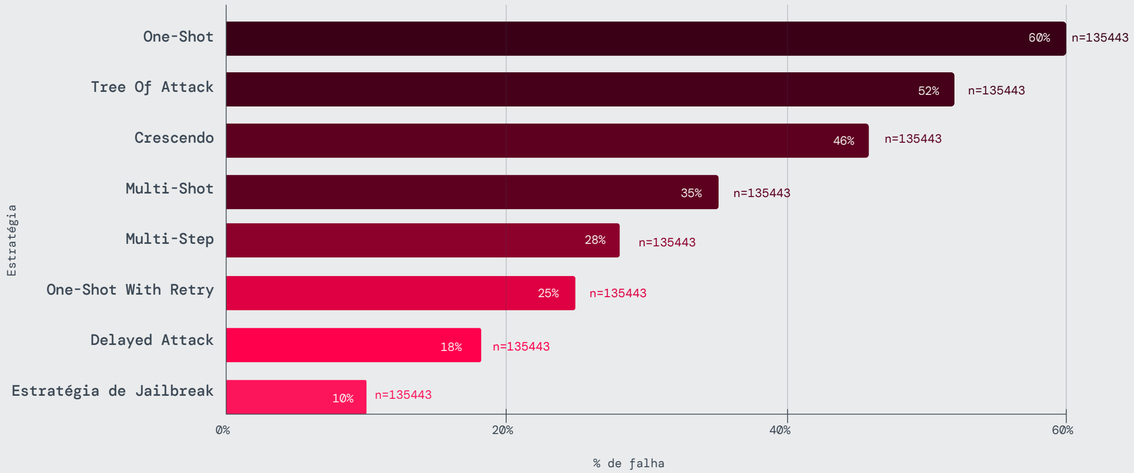



Figura 19: distribuição das principais variações (técnicas de exploit que modificam as entradas) por taxa de falha. Apenas os tipos de variação com ≥ 50 tentativas estão incluídos.

### O QUE ISSO SIGNIFICA PARA AS EQUIPES DE SEGURANÇA

Esse estudo de caso demonstra que o risco da IA empresarial é inerente e persistente. As falhas surgem repetidamente em áreas de risco conhecidas e quase imediatamente após os sistemas serem testados. Sem testes e controles contínuos, os sistemas de IA introduzem riscos materiais desde o momento em que os modelos são implementados.

# A mais nova fase_ da governança de IA

Em 2025, o foco se expandiu dos princípios éticos e de como a IA deveria se comportar para a segurança com que ela deve operar. Com isso, surgiram novas exigências para controles de risco, testes e supervisão contínua em todo o mundo.

## A segurança no centro da lei de IA da União Europeia em meio a prazos em mudança

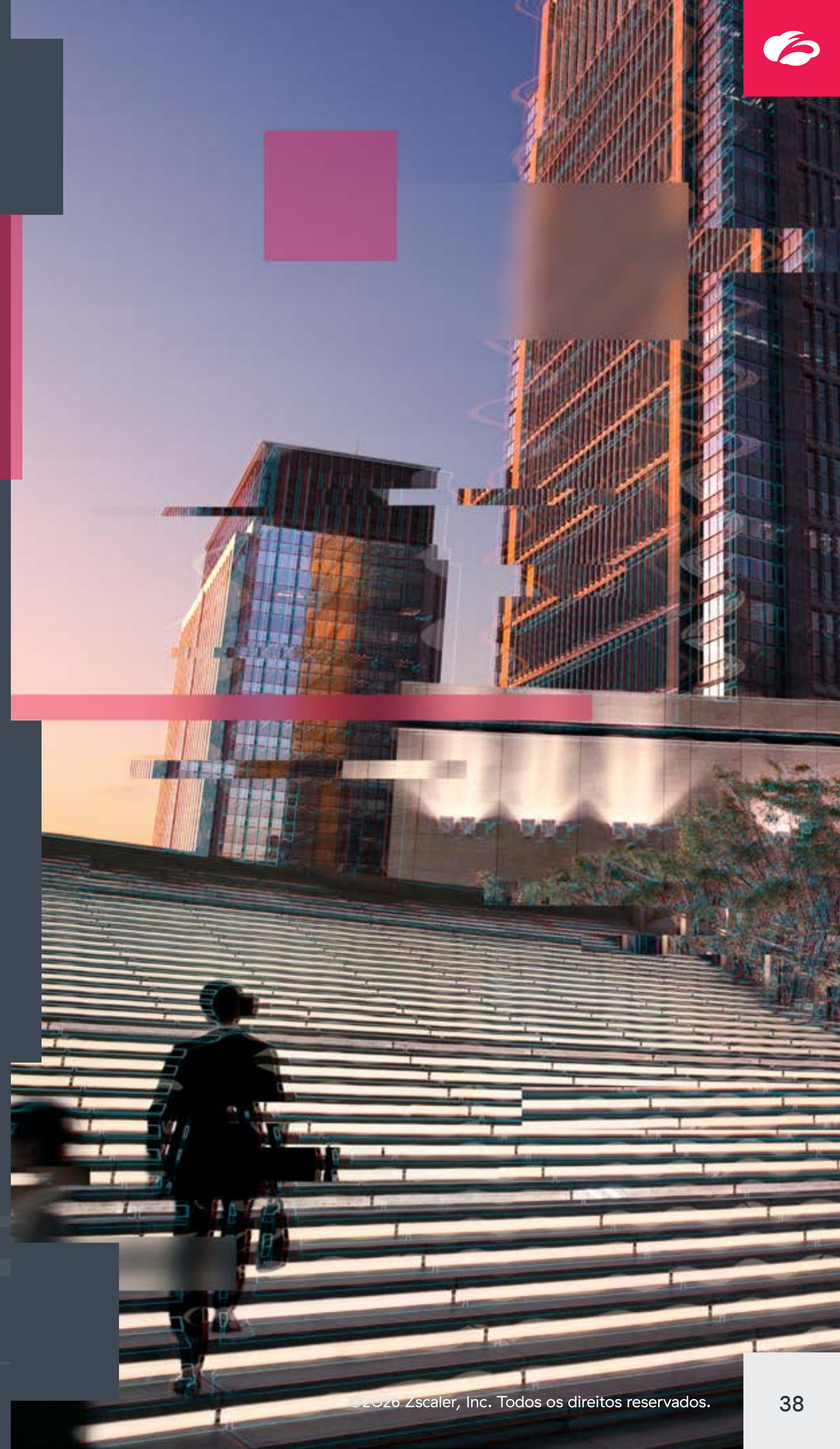
A lei de inteligência artificial da União Europeia continua sendo a estrutura regulatória de IA mais abrangente, mas os cronogramas de implementação e as expectativas de fiscalização estão em constante mudança. No final de 2025, a Comissão Europeia propôs estender os prazos de conformidade para as partes mais arriscadas da lei, particularmente os sistemas de IA de alto risco (usados em saúde, segurança pública, etc.), até dezembro de 2027, mediante aprovação do parlamento e dos Estados-Membros.[3] Ao mesmo tempo, novas orientações e plataformas de suporte estão sendo implementadas para ajudar as organizações a lidar com requisitos como notificação de incidentes e avaliações de conformidade.[4]

As organizações devem encarar a lei de IA da UE não como um prazo de conformidade estático, mas como um alvo móvel, que exige prontidão contínua e controles de segurança proativos.

## A governança da IA nos EUA baseia-se em padrões, não em leis

Os Estados Unidos ainda não possuem uma legislação federal abrangente sobre IA, mas 2025 marcou uma clara mudança na forma como o governo americano encara a IA: priorizando a competitividade nacional, com a segurança e a governança pautadas por padrões e políticas de agências, em vez de uma regulamentação ampla. O Instituto Nacional de Padrões e Tecnologia (NIST) continua liderando a adoção da AI Risk Management Framework[5] como base para o desenvolvimento seguro, testes de adversários e garantias operacionais.

Em dezembro de 2025, o governo emitiu uma ordem executiva com o objetivo de impedir ou contestar leis estaduais de IA que conflitem com uma estrutura política nacional de IA e orientar as agências a buscar padrões federais e recorrer a litígios quando necessário.[6] Apesar disso, vários estados (incluindo Nova York)[7] continuam a avançar com suas próprias leis de segurança de IA, ressaltando que a regulamentação de IA nos EUA em 2026 envolverá a navegação em um complexo ambiente político federal-estadual.



[3] Reuters, EU to delay 'high risk' AI rules until 2027 after Big Tech pushback, 19 de novembro de 2025.

[4] Comissão Europeia, Commission launches AI Act Service Desk and Single Information Platform to support AI Act implementation, 8 de outubro de 2025.

[5] NIST, AI Risk Management Framesita.

[6] Axios, Executive order targeting state AI laws, 11 de dezembro de 2025.

[7] Axios, N.Y. Gov. Kathy Hochul signs sweeping AI safety bill, 19 de dezembro de 2025.

# A região da APAC acelera a adoção segura da IA

Em toda a região da Ásia-Pacífico, os governos continuam a promover estratégias de IA que vinculam explicitamente a rápida adoção à segurança e à resiliência. Muitas economias da região da Ásia-Pacífico estão enfatizando estruturas de governança práticas e controles baseados em risco que possam ser dimensionados juntamente com a implementação da IA.

O Japão deu um passo importante em 2025 com a aprovação de sua primeira lei abrangente sobre IA, a Lei de Promoção da IA,[8] em maio de 2025, estabelecendo um plano nacional que promove a pesquisa, o desenvolvimento em e a implantação de IA, reconhecendo formalmente a necessidade de gerenciar os riscos associados.

A Índia seguiu o exemplo com suas Diretrizes de Governança de IA de 2025,[9] uma estrutura ampla voltada para uma "IA segura e confiável". Essas diretrizes vinculam a adoção de IA à infraestrutura pública digital do país e estabelecem expectativas para governança de dados, transparência algorítmica e gerenciamento de riscos, particularmente para serviços públicos de grande escala e sistemas financeiros.

Singapura continuou a aprimorar seu ecossistema de governança de IA até 2025, expandindo sua estrutura de testes AI Verify e iniciativas de garantia de GenAI relacionadas,[10] avançando ainda mais em direção a testes, monitoramento e garantia contínuos.

A Austrália também avançou em sua abordagem por meio do guia para adoção de IA, lançado em outubro de 2025,[11] juntamente com sua agenda de IA segura e responsável; esforços que enfatizam proteções, testes e maior supervisão para implantações de maior risco, particularmente em setores regulamentados.

Com diversos planos estratégicos substanciais para 2025 avançando em paralelo, a região da Ásia-Pacífico está se posicionando cada vez mais como líder global em inovação e adoção de IA pragmática e com foco em segurança.

As expectativas em relação à segurança de IA devem aumentar acentuadamente em 2026. Mesmo com a evolução da governança global e regional (e a aplicação ainda desigual), as organizações precisarão assumir a responsabilidade por garantir a adoção da IA. Os formuladores de políticas podem pressionar por controles baseados em evidências, mas a convergência de estruturas por si só não reduzirá o risco. O sucesso da IA dependerá, em última análise, da disciplina de segurança interna. Organizações que implementam zero trust, testam continuamente seus modelos e monitoram a evolução das ameaças estarão em melhor posição para implantar a IA de forma responsável.

[8] IT Business Today, **Japan's AI Regulation is a Significant Step Forward with the AI Promotion Act,** 29 de outubro de 2025.

[9] AI, Data & Analytics Network, **India unveils new AI governance guidelines to encourage responsible adoption,** 6 de novembro de 2025.

[10] IMDA, **Singapore launches new tools to help businesses protect data and deploy AI in a trusted ecosystem,** 7 de julho de 2025.

[11] Governo australiano, DISR, **Guidance for AI Adoption,** 21 de outubro de 2025.

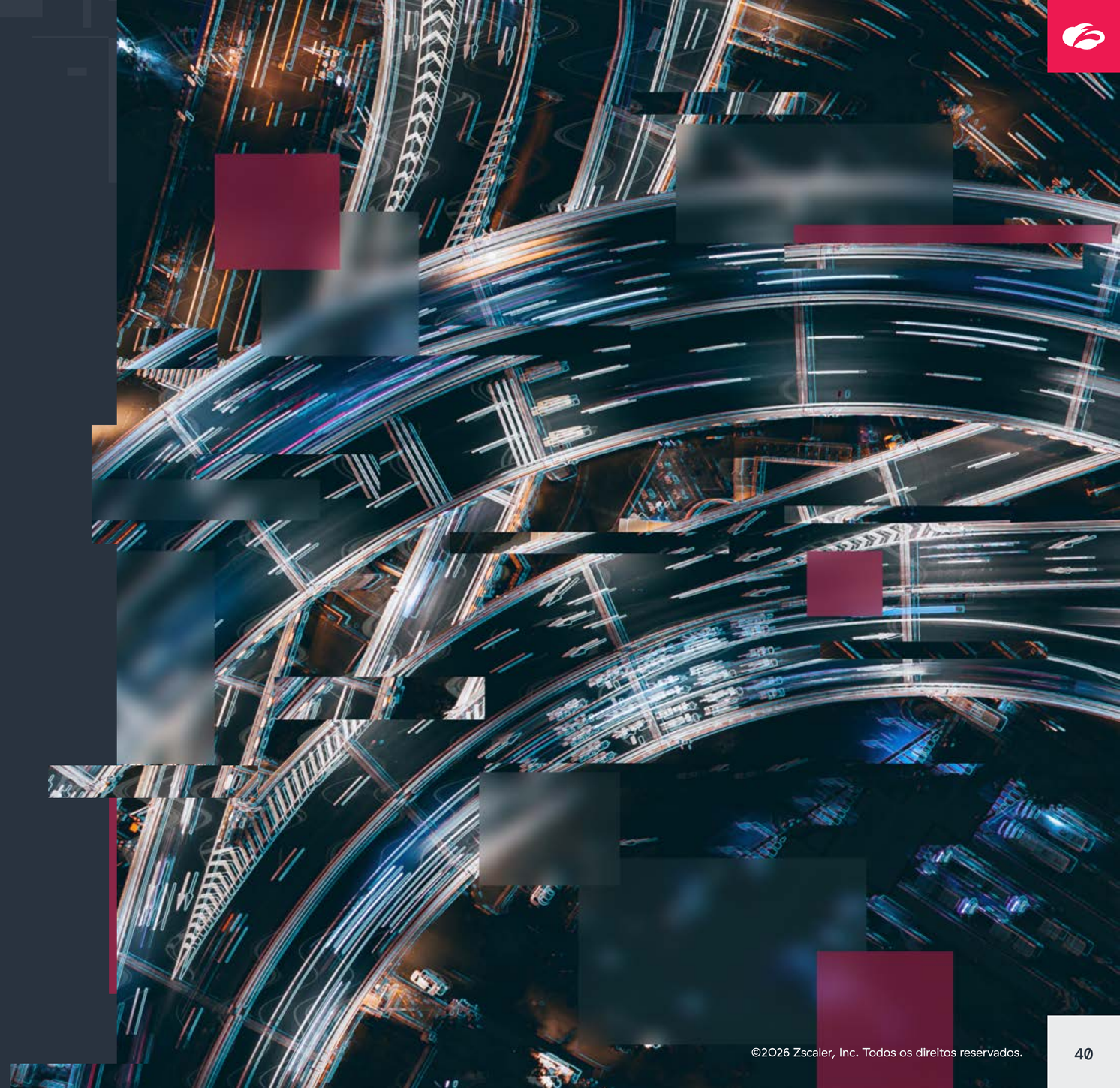
# Previsões de segurança de IA para 2026

## 1 Ataques de IA agêntica autônomos e orquestrados por humanos

A ameaça da IA agêntica aumentará à medida que os sistemas autônomos assumirem uma parcela maior da carga de trabalho de intrusão. Agentes de IA capazes de planejar e executar ações de forma independente desempenharão um papel mais importante nos ataques cibernéticos em 2026. Os primeiros sinais dessa mudança já apareceram em 2025 com a **primeira campanha de espionagem orquestrada por IA relatada**, como mencionado anteriormente, na qual um grupo patrocinado por um governo automatizou de 80% a 90% das etapas de seu ataque com IA agêntica. Os ataques de ransomware auxiliados por IA acelerarão a transição da criptografia para o roubo de dados em alta velocidade, com a IA permitindo realizar mais operações simultaneamente e reduzindo os custos operacionais do atacante.

## 2 Ataques à cadeia de suprimentos de IA

Os ataques à cadeia de suprimentos de IA terão como alvo os componentes essenciais que alimentam os sistemas de IA corporativos. **As descobertas da ThreatLabz** em 2025 expuseram como vulnerabilidades em arquivos de modelos e camadas de processamento comuns podem ser exploradas para acessar sistemas sigilosos. Os invasores se concentrarão cada vez mais em adulterar os componentes subjacentes da IA (modelos e conjuntos de dados), em vez de apenas utilizá-la indevidamente no nível do aplicativo. À medida que mais organizações importam componentes de IA de terceiros para seus ambientes, comprometer esses elementos fundamentais proporcionará acesso poderoso. Proteger a cadeia de suprimentos de IA continuará sendo tão importante quanto proteger o aplicativo construído sobre ela.



## 3 Riscos de segurança da IA embarcada

A inteligência artificial embarcada em aplicativos do dia a dia introduzirá acessos ocultos que as ferramentas de segurança tradicionais podem não perceber. Recursos de IA integrados diretamente em aplicativos comerciais populares, plataformas em nuvem e ferramentas móveis (como os resumos de reuniões com IA do Zoom ou o assistente Copilot do Microsoft 365) criarão riscos sutis que são fáceis de passar despercebidos. Esses recursos de IA embarcados geralmente têm amplo acesso a conteúdo sigilosos, tornando-as alvos atraentes para uso indevido. As empresas devem esperar que os atacantes tentem cada vez mais explorar essas funções integradas para exfiltrar informações valiosas ou obter acesso e se movimentar silenciosamente dentro de um ambiente, aproveitando-se do fato de que muitas organizações ainda não têm visibilidade completa de onde a IA foi incorporada na cadeia de suprimentos de software.

## 4 Ransomware e ataques patrocinados por governos a repositórios de dados de GenAI

À medida que as empresas passam dos projetos-piloto de GenAI para implementações completas em 2026, um número muito maior de sistemas internos encaminhará informações sigilosas para fluxos de trabalho orientados por IA. Os atacantes aproveitarão essa mudança visando os bancos de dados por trás dos aplicativos de GenAI. Esses repositórios contêm mais do que dados brutos, mas também contexto e intenção, dando aos adversários uma visibilidade muito maior dos ciclos de decisão internos e, como resultado, mais poder de influência do que a maioria das violações tradicionais oferece. Comprometer os repositórios de dados de LLMs se tornará uma tática altamente eficaz para espionagem e extorsão por meio de ransomware no próximo ano.

## 5 IA fraudulenta incorporada aos fluxos de trabalho corporativos

Serviços e plataformas de IA fraudulentos deixarão de ser golpes isolados para se tornarem uma presença constante e integrada aos fluxos de trabalho empresariais. O crescimento constante da adoção de ferramentas de IA em 2025 já demonstrou a facilidade com que serviços maliciosos de IA podem se infiltrar em fluxos de trabalho reais. A expectativa é que os atacantes evoluam de páginas de IA fraudulentas para o lançamento de copilotos maliciosos completos, capazes de se comportar como assistentes de produtividade reais e se integrar de forma discreta ao uso diário. Essa próxima fase tornará os assistentes não autorizados mais difíceis de detectar, contribuindo significativamente para os riscos da IA não aprovada ou paralela usada por funcionários de empresas.

## 6 Segurança e responsabilização de IA em toda a organização

A segurança de IA se tornará um requisito para toda a empresa à medida que a supervisão e a responsabilidade aumentarem. Após um ano de grandes preocupações e crescente escrutínio em 2025, as organizações enfrentam expectativas cada vez maiores sobre como gerenciar a IA: como os modelos são avaliados, como os dados são tratados e como o potencial uso indevido é monitorado. Em 2026, garantir a segurança dos sistemas de IA não será mais opcional nem se limitará às equipes técnicas. Os líderes precisarão ter uma visão clara dos riscos da IA, e as políticas de segurança precisam abranger todas as áreas da empresa que interagem com IA.

# Práticas_recomendadas: adoção segura de IA nas empresas

## 5 verdades duras sobre a segurança de IA em 2026

1. **Você não pode proteger o que não pode ver. A IA paralela e a funcionalidade de IA embarcada fazem da visibilidade o novo perímetro.**

2. **As configurações padrão do fornecedor não são projetadas para lidar com riscos corporativos. Recursos de IA frequentemente são disponibilizados já ativados e com permissões excessivamente amplas.**

3. **A governança da IA é um alvo em constante movimento. As políticas devem evoluir à medida que os recursos e as ameaças mudam.**

4. **O princípio de zero trust agora se estende aos modelos de IA. Eles exigem o mesmo nível de controle de acesso que os usuários humanos.**

5. **A inteligência artificial é, inegavelmente, uma parte da superfície de ataque. Vulnerabilidades em modelos e ataques com IA agêntica já estão ocorrendo.**

A boa notícia: você não precisa aceitar essas "verdades duras" como o preço da adoção da IA. Use a lista de verificação de segurança empresarial de 2026 a seguir para priorizar as proteções adequadas.

# Lista de verificação de segurança de IA empresarial para 2026

As seguintes práticas recomendadas estabelecem uma base sólida para o uso seguro da IA.

**Inventarie todos os aplicativos de GenAI e aplicativos com funcionalidade de IA integrada**

- [ ] Crie um catálogo continuamente atualizado de todas as ferramentas de GenAI independentes e de todos os aplicativos SaaS ou internos que incluam funcionalidades ou recursos de IA.

**Desative configurações padrão de IA que representem risco**

- [ ] Desative a funcionalidade de IA ativada automaticamente em aplicativos SaaS e de produtividade até que sejam revisados e configurados para corresponder ao seu perfil de risco.

**Aplique zero trust a todas as interações do modelo**

- [ ] Implemente o princípio de privilégio mínimo para todos os usuários, serviços e sistemas que interagem com um modelo de IA.

**Aplique medidas de segurança de IA com inspeção inline**

- [ ] Garanta a inspeção inline de todo o tráfego de IA/ML para impedir que atividades maliciosas externas comprometam os sistemas de IA e evitar que dados sigilosos sejam expostos por meio de prompts ou em resultados.

**Valide a linhagem do modelo e a cadeia de suprimentos**

- [ ] Verifique a procedência, as atualizações, os conjuntos de dados e as dependências de cada modelo para reduzir o risco de adulteração, contaminação ou componentes comprometidos.

As empresas também devem definir padrões de governança e regras de engajamento para a forma como a IA é adotada e gerenciada.

**Atualize a governança de IA com frequência**

- [ ] Atualize regularmente as políticas, os controles de acesso e as classificações de risco para acompanhar as rápidas mudanças nos recursos de IA e nos requisitos regulatórios.

**Exija revisão humana para fluxos de trabalho regulamentados**

- [ ] Assegure-se de que pessoas permaneçam envolvidas sempre que a IA influenciar decisões relacionadas à segurança, conformidade, decisões financeiras ou determinações do setor público.

**Realize testes adversários e simule situações de red teaming**

- [ ] Teste continuamente os modelos em busca de vulnerabilidades como jailbreaks, injeção de código, vazamento de dados e outras fragilidades exploráveis antes que os atacantes as encontrem.

**Garanta a segurança do ciclo de vida de desenvolvimento de IA de ponta a ponta**

- [ ] Aplique controles de assimilação do conjunto de dados, treinamento, implantação e monitoramento para evitar que vulnerabilidades entrem nos sistemas de produção.

# Como as empresas estão implementando a GenAI com segurança: um guia prático

Em 2025, os riscos da IA surgiram de ambos os lados da fronteira empresarial. Os agentes maliciosos usaram GenAI para acelerar e facilitar suas operações, enquanto a exposição interna decorria cada vez mais do uso cotidiano da IA sem supervisão formal, permitindo que os dados chegassem aos sistemas de IA antes que as equipes de segurança pudessem avaliar ou controlar o risco.

As organizações que evitaram incidentes foram aquelas que introduziram a GenAI em fases controladas e habilitaram apenas o que podiam governar.

**O plano de ação deles no mundo real é o seguinte:**

### COMECE COM UMA POSTURA DE ZERO TRUST E RESTRINJA OS SERVIÇOS DE IA NÃO VERIFICADOS

Inúmeras ferramentas de IA introduzem riscos desconhecidos de segurança e tratamento de dados, tornando crucial partir de uma posição de zero trust. Bloquear ou limitar o acesso a aplicativos de IA/ML não avaliados elimina a exposição imediata e previne vazamentos iniciais de dados, oferecendo às equipes de segurança o tempo necessário para avaliar quais aplicativos são apropriados para uso corporativo.

### HOSPEDE FERRAMENTAS DE GENAI APROVADAS EM UM AMBIENTE PRIVADO E CONTROLADO

Para manter o controle total sobre os dados corporativos, as organizações devem executar as ferramentas de GenAI aprovadas em um ambiente privado e seguro, como uma instância dedicada ou uma isolada, gerenciada inteiramente pela empresa. Essa configuração garante que nem o fornecedor nem terceiros possam acessar dados internos ou de clientes e impede que prompts e resultados sejam usados para treinar modelos públicos. Operar a GenAI dessa forma preserva a soberania dos dados e impede que informações sigilosas saiam da organização.

### IDENTIFIQUE E VALIDE OS APLICATIVOS DE GENAI QUE ATENDEM AOS REQUISITOS DA EMPRESA

Determine quais aplicativos de GenAI são seguros para usar, verificando como eles lidam com os dados, se mantêm suas informações isoladas, como o modelo foi construído e se o fornecedor atende aos seus requisitos de segurança, privacidade e conformidade. Somente as ferramentas que atenderem a esses padrões devem prosseguir.

### IMPLEMENTE CONTROLES DE IDENTIDADE E ACESSO RIGOROSOS

Posicione os aplicativos de GenAI aprovados atrás de uma arquitetura zero trust com políticas de acesso granulares. Isso garante que cada usuário, departamento e fluxo de trabalho receba apenas o acesso necessário, ao mesmo tempo que oferece às equipes de segurança visibilidade e controle de ponta a ponta sobre todas as atividades.

### APLIQUE MEDIDAS DE PROTEÇÃO DE DADOS PARA EVITAR O COMPARTILHAMENTO ACIDENTAL OU NÃO AUTORIZADO

Combine o acesso aprovado com DLP de nível empresarial. O monitoramento e a inspeção do tráfego de e para aplicativos de IA garantem que as informações sigilosas permaneçam protegidas e que nenhum dado crítico seja exposto por meio de interações com esses aplicativos.

# Como a Zscaler oferece_ proteção de IA abrangente

As conclusões deste relatório confirmam que a adoção da IA nas empresas está acelerarando rapidamente. Como resultado, uma superfície de ataque crescente, o uso de IA paralela e embarcada, e modelos e infraestrutura em constante evolução estão introduzindo novos riscos relacionados à exposição, uso indevido e governança de dados, que as abordagens de segurança tradicionais não conseguem abordar de forma eficaz.

As arquiteturas de segurança baseadas em firewalls, VPNs e controles de perímetro não foram projetadas nem destinadas a ambientes dinâmicos de IA. Na prática, elas aumentam a complexidade e deixam falhas de visibilidade. Elas têm dificuldade em aplicar controles consistentes em ferramentas públicas de IA, agentes, modelos privados e componentes emergentes, como servidores de Model Context Protocol (MCP).

As organizações acabam reagindo aos riscos da IA em vez de gerenciá-los proativamente.

Garantir a segurança da IA em grande escala exige uma abordagem diferente, que reduza a exposição por padrão, verifique continuamente o acesso e aplique controles de segurança onde quer que a IA seja usada ou implementada. O zero trust fornece essa base.

A Zscaler oferece uma plataforma de segurança de IA baseada em zero trust, que protege a IA em todos os lugares, em todas as formas como as organizações usam, criam e operam a IA. Ao reduzir a superfície de ataque, aplicar o princípio de privilégio mínimo e inspecionar todo o tráfego inline, a Zscaler ajuda as organizações a adotarem a IA de forma segura, sem comprometer a inovação.

# Transformando o risco da IA em adoção segura da IA

Com o zero trust como base, a Zscaler aplica controles de segurança nativos de IA que traduzem a arquitetura em ação. Essas funcionalidades oferecem às organizações a visibilidade, os mecanismos de controle e as proteções necessárias para governar o uso da IA em tempo real, ao mesmo tempo que combatem ativamente as ameaças baseadas em IA para usuários, aplicativos e infraestrutura.

## A Zscaler IA permite que as organizações:

### OFEREÇA O USO SEGURO DE IA EM CONTEXTOS PÚBLICOS E PRIVADOS

- Veja exatamente onde e como a IA está sendo usada, incluindo aplicativos, modelos, agentes, instruções, respostas e componentes emergentes, como servidores de MCP.
- Permita que os funcionários usem ferramentas de IA de forma produtiva, isolando interações de IA baseadas na web que possam ser de risco e impedindo que dados sigilosos sejam compartilhados involuntariamente com modelos externos.
- Detecte e bloqueie injeção de prompts, exposição de informações pessoais identificáveis (PII), envenenamento de dados, saídas inseguras e outras ameaças específicas de IA em tempo de execução com proteções de IA integradas.
- Controle quem pode usar IA, a quais ferramentas os usuários podem ter acesso e como a IA é usada com políticas que se adaptam continuamente ao risco do usuário, do dispositivo e do aplicativo, bloqueando automaticamente a IA não autorizada ou paralela.
- Impeça o envio ou o recebimento de dados sigilosos por ferramentas de IA usando controles de DLP integrados e com reconhecimento de IA.
- Mantenha um registro de auditoria detalhado e pesquisável das atividades de IA para dar suporte a investigações e conformidade.

### MANTENHA-SE À FRENTE DAS AMEAÇAS BASEADAS EM IA

- Reduza a exposição eliminando a superfície de ataque externa e aplicando verificação contínua e acesso de privilégio mínimo.
- Inspecione todo o tráfego, incluindo tráfego criptografado, para bloquear ameaças aprimoradas por IA em tempo real.
- Aplique IA preditiva e generativa para identificar riscos mais rapidamente e melhorar as operações e a resposta de segurança.
- Descubra, classifique e proteja continuamente dados sigilosos em terminais, tráfego inline e ambientes em nuvem.
- Impeça a movimentação lateral com segmentação baseada em IA, que limita o alcance do ataque.
- Avalie continuamente a IA e a postura de zero trust com insights e recomendações gerados por IA.

Esses resultados são alcançados por meio de um conjunto unificado de proteções que abrangem todo o ciclo de vida da segurança da IA, conforme abordado na seção a seguir.

# Zscaler + IA: protegendo a forma como as organizações usam e criam aplicativos

A Zscaler oferece proteção abrangente, desde a descoberta e avaliação de riscos até a segurança de aplicativos e acessos de IA, abrangendo IA pública e privada, modelos, pipelines, agentes e infraestrutura.

## GERENCIAMENTO DE ATIVOS DE IA

### Descubra todo o seu uso de IA e os riscos envolvidos

- **Visibilidade completa** de todos os aplicativos, modelos, pipelines e servidores de MCP.
- **Um AI-BOM** para identificar riscos na cadeia de suprimentos e de dependências.
- **Identificação de** aplicativos SaaS de GenAI e modelos de IA.

## PROTEJA O ACESSO A APLICATIVOS DE IA

### Garanta o uso seguro e responsável de aplicativos de IA

- **Controle granular** sobre quais usuários podem acessar quais aplicativos.
- **Inspeção inline** de prompts e respostas para evitar o envio ou retorno de dados sigilosos.
- **Controles de conteúdo** para bloquear saídas inseguras ou prejudiciais.

## PROTEJA APLICATIVOS E INFRAESTRUTURA DE IA

### Reforce os sistemas e prompts de IA e aplique proteção em tempo de execução

- **Detecção de vulnerabilidades** em modelos e pipelines.
- **Testes de red team** para identificar vulnerabilidades e pontos fracos.
- **Proteção contra injeções de prompt**, envenenamento de dados, uso de dados sigilosos, etc.

**Governança de IA:** mantenha-se em conformidade com as estruturas de IA através do mapeamento dos controles de segurança de IA para a Estrutura de Gestão de Riscos de IA do NIST e a Lei de IA da UE.

# Metodologia_ de pesquisa

As conclusões são baseadas na análise de 989,3 bilhões de transações totais de IA e ML na nuvem da Zscaler, de janeiro a dezembro de 2025. A nuvem de segurança global da Zscaler processa mais de 500 trilhões de sinais diários e bloqueia mais de 9 bilhões de ameaças e violações de políticas por dia, fornecendo mais de 250 mil atualizações de segurança diárias.

# Sobre a_ ThreatLabz

A ThreatLabz é o braço de pesquisa de segurança da Zscaler. Essa equipe de alto nível é responsável por identificar novas ameaças e garantir que as milhares de organizações que utilizam a plataforma global da Zscaler estejam sempre protegidas. Além da pesquisa de malware e da análise comportamental, os membros da equipe estão envolvidos na pesquisa e no desenvolvimento de novos módulos protótipos para proteção avançada contra ameaças na plataforma da Zscaler e realizam regularmente auditorias internas de segurança para garantir que os produtos e a infraestrutura da Zscaler atendam aos padrões de conformidade de segurança. A ThreatLabz publica com frequência análises aprofundadas de ameaças novas e emergentes em seu portal, **research.zscaler.com**.

Siganos: X **@ThreatLabz** | ThreatLabz **security research blog**

Zero Trust Everywhere

**Sobre a Zscaler**

A Zscaler (NASDAQ: ZS) acelera a transformação digital para que seus clientes possam ter mais agilidade, eficiência, resiliência e segurança. A Zscaler Zero Trust Exchange™ protege milhares de clientes contra ataques cibernéticos e perda de dados, conectando com segurança usuários, dispositivos e aplicativos em qualquer local. Distribuída em mais de 150 data centers globalmente, a Zero Trust Exchange™ baseada em SSE é a maior plataforma integrada de segurança na nuvem do mundo. Saiba mais em **zscaler.com/br** ou siga-nos no Twitter **@zscaler**.



**+1 408.533.0288** **Zscaler, Inc. (Sede) • 120 Holger Way • San Jose, CA 95134** **zscaler.com/br**