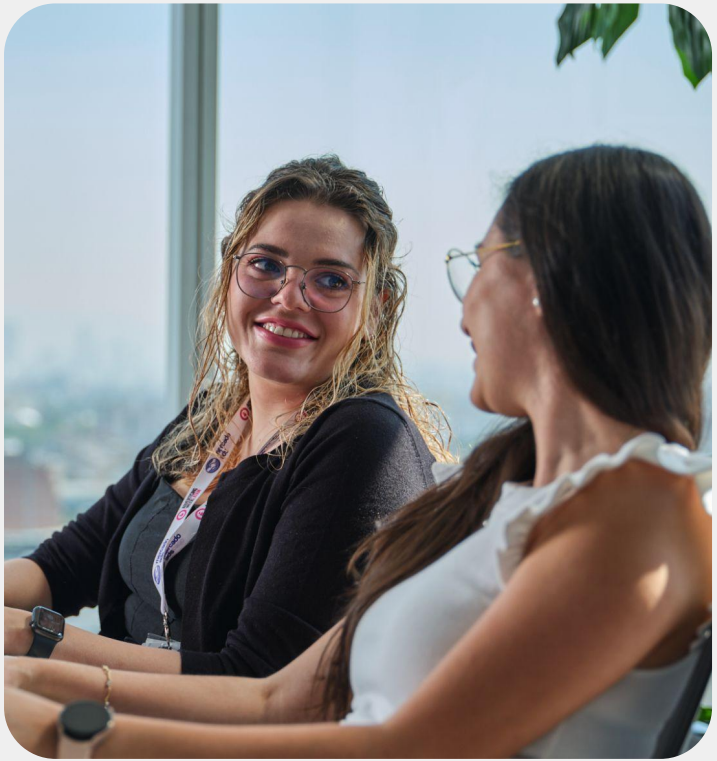
Segurança e Privacidade em LLMs e Agentes

GEN AI

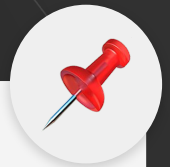
Neocamp

NeoCamp

Segurança e Privacidade em LLMs e Agentes



Segurança e Privacidade em LLMs e Agentes



A **IA generativa**, como os **LLMs** e **Agentes**, oferece grandes oportunidades

No entanto, seu rápido crescimento apresenta novos riscos em **segurança** e **privacidade**

Esta apresentação oferece práticas para um uso **seguro** e **responsável**

O que são LLMs e Agentes?

LLMs (Modelos de Linguagem de Grande Porte) são sistemas de IA treinados com enormes quantidades de texto para entender e gerar linguagem natural, como responder perguntas ou criar textos
Exemplo: ChatGPT

Agente é um sistema que percebe o ambiente e age sobre ele para cumprir tarefas. Pode depender de comandos ou supervisão humana para funcionar
Exemplo: Assistente Virtual de Smartphone

Agente Autônomo é um agente que toma decisões e age sozinho, sem precisar de supervisão contínua, usando sua própria “inteligência” para decidir o que fazer
Exemplo: Robô de entrega autônoma

Riscos de Segurança em LLMs

01 02 03

Ataques de Injeção

Manipulação da saída do modelo;
O objetivo é revelar informações.

Vulnerabilidades de Integração

Fraquezas ao conectar LLMs com outros sistemas;
Abertura de portas para ataques.

Exfiltração de Dados

Roubo de informações sensíveis por meio de prompts maliciosos;
Um risco constante.

Exemplos incluem a injeção de prompts para obter chaves de API ou ataques de negação de serviço (DoS - *Denial of Service*) com prompts excessivos

Riscos de Privacidade em LLMs

01 02 03

Coleta de Dados

Os prompts podem conter informações pessoais;
Precisamos ser cuidadosos.

Uso Indevido de Dados

Treinamento com dados sensíveis
Inferência de informações privada.

Divulgação Acidental

Respostas de LLMs podem filtrar dados confidenciais;
Um risco.

Isto inclui inferir dados demográficos da linguagem ou reter prompts com PII (*Informações Pessoalmente Identificáveis*) indefinidamente

Práticas de Segurança Recomendadas

➔ **Validar Prompts**

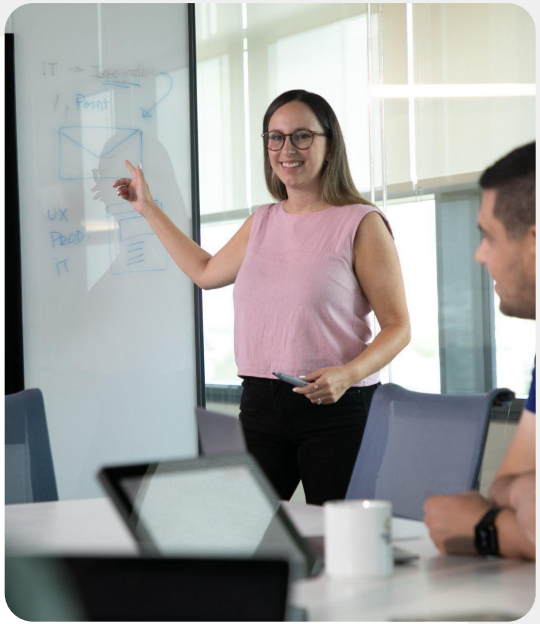
Sanitizar Prompts antes de enviar para o LLM
Previne injeções maliciosas

➔ **Controles de Acesso**

Implementar autenticação robusta
Somente usuários autorizados têm acesso

➔ **Monitorar o Uso**

Auditar LLMs para detectar anomalias
Identifica comportamentos suspeitos



Recomenda-se usar listas brancas de palavras-chave e limite de taxa para prevenir ataques DoS - *Denial of Service*

Práticas de Privacidade Recomendadas

➔ **Anonimizar Dados**

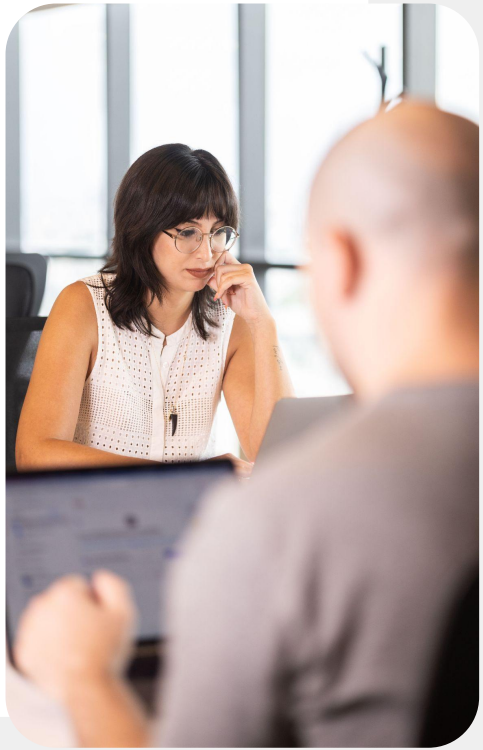
Remover informações pessoais dos prompts;
Proteja a identidade do usuário.

➔ **Informar os Usuários**

Transparência sobre o uso de seus dados;
Gera confiança e cumprimento.

➔ **Cumprir Regulamentações**

Adotar o LGPD - *Lei Geral de Proteção de Dados Pessoais*, GDPR - *General Data Protection Regulation* e outros;
Evita multas e protege a privacidade;
Isso inclui técnicas de mascaramento ou generalização de dados, e políticas claras de retenção.



Segurança e Privacidade em Agentes Autônomos

1. Mínimo Privilégio

Limitar o acesso a recursos;

Somente o necessário para a tarefa.

2. Zero Trust

Não confie em nada;

Verifique sempre cada interação.

3. Auditar Ações

Monitorar em tempo real;

Detectar comportamentos incomuns.

É crucial restringir o acesso do agente e manter um registro detalhado de todas as suas ações

Ferramentas e Técnicas

Bibliotecas de Segurança	Deepchecks, Arthur AI	Detecta vieses e vulnerabilidades nos modelos LLM
Técnicas de Privacidade	Privacidade Diferencial	Protege a confidencialidade no treinamento do LLM
Plataformas de Monitoramento	Análise em tempo real	Monitora a segurança do LLM constantemente

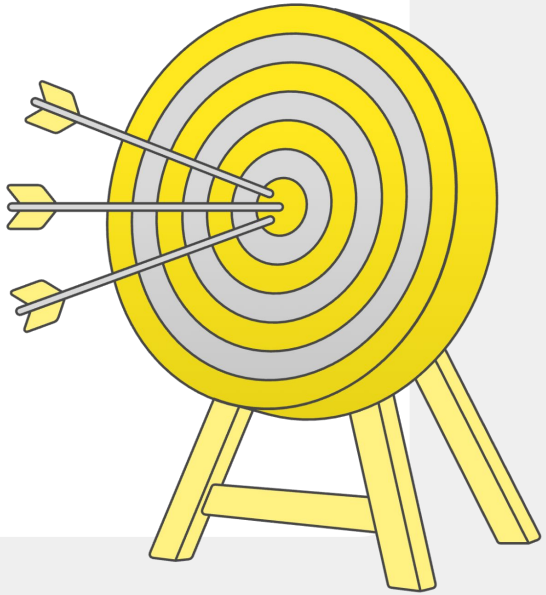
Essas ferramentas ajudam a fortalecer a postura de segurança e privacidade no desenvolvimento e uso da IA

Conclusão

A **segurança** e a **privacidade** são os pilares do uso responsável dos LLMs

A implementação de **boas práticas** reduz os riscos e gera confiança

A colaboração entre especialistas em segurança, privacidade e IA é vital para a construção de um futuro **seguro** e **ético**



Dúvidas?

Obrigado pela presença!

Emojis