



Inteligência Artificial



Produção:

cert.br nic.br cgi.br

IA É ASSISTENTE. O RESPONSÁVEL É VOCÊ!

Uso da Inteligência Artificial (IA) já faz parte do dia a dia. Ela ajuda a escrever textos, criar imagens, fazer traduções e até executar tarefas. No entanto, não dá para confiar em tudo o que a IA faz. Os resultados podem conter erros, expor dados ou ser usados para golpes.

Veja aqui como tirar proveito das ferramentas de IA de forma segura e responsável.



*Tirou as informações
sensíveis do prompt?*



Checou a resposta?



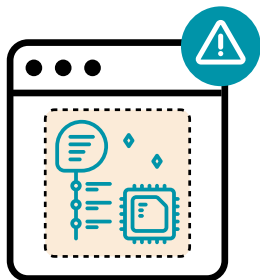
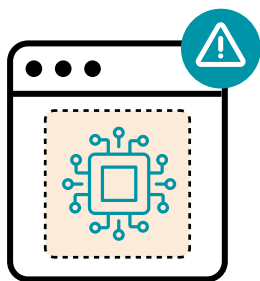
*Limitou as ações dos
agentes?*

***A RESPONSABILIDADE
É SUA***

ENTENDA AS LIMITAÇÕES DA IA

Embora seja comum dizer “a IA”, não existe uma IA geral que faça tudo. **São vários tipos de IA**, com fins específicos, e **todos têm limitações**. Grandes modelos de linguagem (LLMs), a tecnologia de IA mais usada atualmente, têm como **principais limitações**:

- » Não “sabe” a resposta certa
 - calcula a **resposta mais provável**
- » Depende dos dados de treinamento
 - dados ruins geram respostas ruins
- » É programada para **sempre dar uma resposta**
 - **mesmo que seja errada**

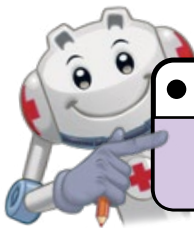


AJUSTE CONFIGURAÇÕES DE PRIVACIDADE

Suas interações com a IA podem ser usadas para melhorar o modelo.

Se você não quiser que seus dados sejam compartilhados, é importante entender e ajustar as opções das ferramentas.

- » **Desative o uso** de suas interações para melhorar o modelo
- » **Desligue o histórico** de atividades
 - principalmente se a ferramenta não oferecer a opção acima
- » Use *chat* temporário ou modo incógnito, se disponível



Veja mais dicas no fascículo
"Privacidade".



NÃO INCLUA DADOS SENSÍVEIS NO PROMPT

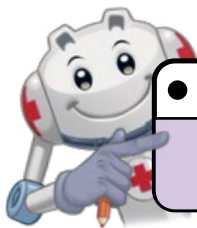
O que você insere no **chatbot** geralmente vai para a empresa que mantém a IA. Seus dados podem ser usados para treinar o modelo, correndo o **risco de aparecer em respostas para outras pessoas**.

» **Assuma que o que vai para a IA é público**

- pergunte-se: “Eu colocaria essa informação em um cartaz na rua?”

» Troque dados sensíveis por termos genéricos, como:

- “nome”, “CPF”, “senha”, “empresa”, “cliente”



Veja mais dicas no fascículo
“Proteção de Dados”.

ELABORE PROMPTS COM CONTEXTO

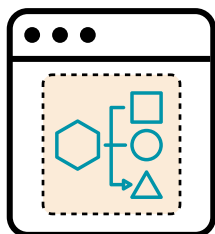
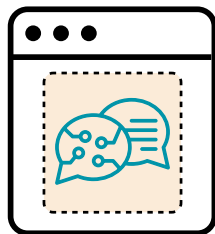
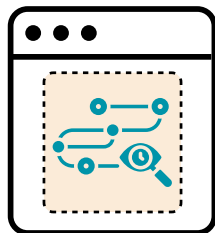
Chatbots são, em essência, geradores de texto e **precisam de contexto** para produzir a resposta desejada. Sem instruções claras e detalhadas, podem gerar respostas erradas ou sem fundamento real (alucinações).

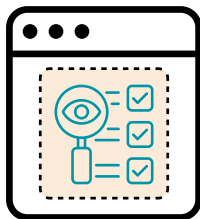
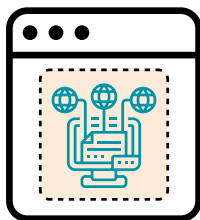
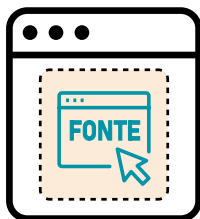
» **Descreva a tarefa** a ser realizada

- indique o que espera como resultado

» **Explique o contexto**, como:

- a situação ou o cenário
- condições a considerar
- quem vai ler (ex: seu chefe, clientes, crianças)
- o tom da linguagem (ex: formal, simples)





FORNEÇA FONTES PARA TEMAS ESPECÍFICOS

Chatbots de uso geral são **treinados com conteúdos variados**. Mas se o *prompt* for sobre um **assunto muito específico**, o risco de respostas inventadas (alucinações) aumenta. Nesses casos, é melhor fornecer uma **fonte especializada** para a IA usar.

- » **Inclua a fonte no *prompt*** e limite a resposta a ela
 - ex: "Com base estritamente na fonte fornecida, responda..."
- » Exija trechos da fonte para validar a resposta
 - ex: "Cite a frase exata da fonte dada que sustenta cada afirmação."



FAÇA A IA RESPONDER QUÊ NÃO SABE

Chatbots são programados para **gerar uma resposta, mesmo quando faltam informações ou há incertezas**. Em vez de responder “não sei”, tentam criar **respostas que podem estar erradas (alucinações)**. Para reduzir esse risco, é necessário dar instruções prévias.

» **Instrua a IA para informar quando não tiver certeza da resposta:**

- ex: “Se faltar informação ou não tiver certeza, diga que não sabe.”



REVISE TODAS AS RESPOSTAS DA IA

Respostas erradas, as chamadas **alucinações**, são um problema conhecido dos LLMs. Usar as respostas de uma IA **sem checar** se estão corretas pode levar a **decisões erradas, perdas financeiras e riscos à saúde**.

- » Trate a resposta da IA como um **rascunho a ser corrigido**
- » Peça à IA que cite as fontes das informações:
 - cheque se são reais e se realmente tratam do assunto
- » **Busque outras fontes e valide:**
 - se os dados e fatos estão corretos
 - se as informações estão atualizadas
- » Lembre-se: **a responsabilidade pelo resultado é sua!**

REMOVA INFORMAÇÕES SENSÍVEIS DA RESPOSTA

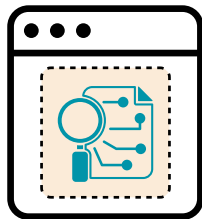
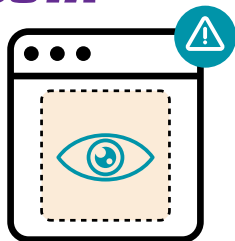
Conteúdos gerados por IA, como imagens, códigos-fonte e documentos, podem conter informações sensíveis, como trechos do *prompt*. Se essas informações não forem removidas, poderão ser expostas indevidamente.

» Examine a resposta além do conteúdo principal

- ex: metadados de arquivos, comentários de código

» Remova as informações que não devem ser divulgadas

- ex: *prompt* original, senhas ou outros dados sensíveis





LIMITE O ACESSO DE AGENTES A INFORMAÇÕES



Salvar senhas, dados de cartão de crédito ou documentos confidenciais **no histórico de interações ou em arquivos** aos quais os agentes têm acesso pode causar **vazamentos de dados e prejuízos financeiros**.

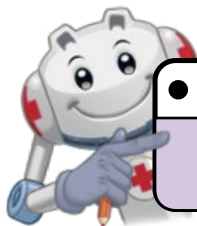
- » Não forneça a agentes informações sensíveis, como senhas e dados bancários
 - **digite as informações** você mesmo, **quando necessário**
- » Defina uma pasta de trabalho para cada agente
 - não grave informações sensíveis nesta pasta



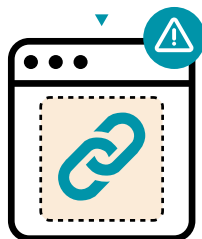
RESTRINJA AS PERMISSÕES DOS AGENTES

Agentes de IA podem ser **manipulados por instruções escondidas** em sites, arquivos ou mensagens (*prompt injection*). Limitar as permissões dos agentes reduz o estrago que um atacante pode causar.

- » **Autorize só o necessário** para realizar a tarefa
 - ex: não permita escrita se o agente só precisa ler
- » Não deixe que o agente acesse todo o sistema
 - defina uma pasta de trabalho



Veja mais dicas no fascículo
"Proteção de Dados".

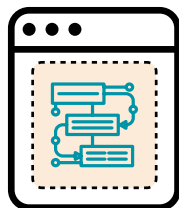


LIMITE O QUE AGENTES Podem DECIDIR SOZINHOS

Agentes de IA podem **interpretar mal instruções ou agir de forma inesperada**. Por exemplo, ao pedir para o agente “limpe meus e-mails”, ele pode apagar todas as mensagens em vez de excluir apenas o *spam*.

» Configure agentes para **exigir sua aprovação** antes de:

- efetivar transações financeiras (ex: pagamentos)
- usar informações sensíveis (ex: senhas, dados bancários)
- enviar mensagens ou publicar conteúdo em seu nome
- executar ações que não podem ser desfeitas (ex: excluir dados)





EVITE O PLÁGIO

Grandes modelos de linguagem são treinados com **imensos volumes de textos** coletados na Internet, **incluindo obras protegidas** por direitos autorais. Ao gerar novos textos, a IA pode reproduzir trechos dessas obras sem citar a fonte, o que pode resultar em plágio.

- » Revise e **reescreva o texto da IA com suas próprias palavras**
- » Cite a fonte e a autoria, sempre que usar trabalhos de outras pessoas
- » Se utilizar IA, indique onde foi usada e com que objetivo
- » Lembre-se: **no final, você é o responsável!**

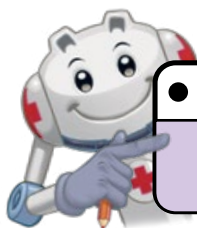
RESPEITE AS REGRAS DA EMPRESA PARA USO DE IA

Pedir para a IA processar **conteúdos que incluem informações internas**, como códigos-fonte de *software*, atas de reuniões ou dados de clientes, pode levar ao **vazamento de informações confidenciais**.

» Antes de usar IA no trabalho, **informe-se** sobre:

- quais ferramentas podem ou não ser usadas
- quais usos são permitidos e proibidos
- como lidar com informações sensíveis

» **Não use contas pessoais** em ferramentas de IA **para tarefas do trabalho**



Veja mais dicas no fascículo
"Trabalho Remoto".

DESCONFIE DE IMAGENS, ÁUDIOS E VÍDEOS

Golpistas usam IA para criar **deepfakes**. Eles fingem ser alguém conhecido ou famoso para pedir dinheiro, fazer propaganda enganosa ou espalhar informações falsas.

» Antes de acreditar, **pergunte-se**:

- essa pessoa realmente diria isso?
- envolve dinheiro, vantagem ou dados pessoais?
- há urgência, ameaça ou apelo emocional?

» Use verificadores de conteúdo feito por IA

- algumas empresas oferecem recursos para verificar conteúdos gerados por suas ferramentas

» Na dúvida, **considere que é falso**



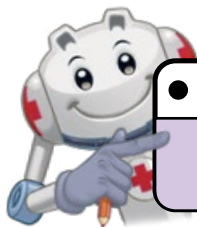
SUSPEITE DE COMPROVANTES DIGITAIS

Golpistas usam ferramentas de IA para **gerar imagens falsas**, como **comprovantes** de transferência e de encomendas despachadas. Com elas, fazem cobranças indevidas, fingem pagamentos ou envio de produtos, para fraudar consumidores.

» Ao fazer compras:

- **faça todo o processo pela plataforma oficial**
- confira o valor e os dados do recebedor antes de concluir o pagamento
- denuncie cobranças recebidas por outros canais

» Ao receber um comprovante de pagamento, **confira o crédito em sua conta**



Veja mais dicas no fascículo
"Golpes: Evite Fraudes".



NÃO TROQUE PROFISSIONAL DE SAÚDE POR IA

Chatbots de uso geral **não têm treino específico** e usam informações nem sempre confiáveis. **Podem dar respostas erradas ou prejudiciais** e gerar pânico ou ilusão de normalidade. Também são incapazes de considerar o contexto do paciente e de atuar em emergências.

» **Não tente se autodiagnosticar ou automedicar** usando IA

- busque atendimento profissional caso não se sinta bem

» **Não faça da IA um terapeuta.** Ela:

- não tem consciência nem emoção e não “entende” você
- tende a concordar, podendo reforçar ideias negativas

» Lembre-se: em interações com IA **não há sigilo profissional**



SAIBA MAIS

- » Para mais detalhes sobre este e outros assuntos relacionados com cuidados na Internet, consulte os demais Fascículos da Cartilha de Segurança para Internet, disponíveis em: **<https://cartilha.cert.br/>**
- » Procurando material para conversar sobre segurança com diferentes públicos? O Portal Internet Segura apresenta uma série de materiais focados em crianças, adolescentes, pais, responsáveis e educadores, confira em: **<https://internetsegura.br/>**

cert.br

O CERT.br (<https://cert.br/>) é um Grupo de Resposta a Incidentes de Segurança (CSIRT) de responsabilidade nacional de último recurso, mantido pelo NIC.br. Além da gestão de incidentes, também atua na conscientização sobre os problemas de segurança, na consciência situacional e transferência de conhecimento, sempre respaldado por forte integração com as comunidades nacional e internacional de CSIRTs.

nic.br

O Núcleo de Informação e Coordenação do Ponto BR - NIC.br (<https://nic.br/>) é uma entidade civil de direito privado e sem fins de lucro, encarregada da operação do domínio .br, bem como da distribuição de números IP e do registro de Sistemas Autônomos no País. Conduz ações e projetos que trazem benefícios à infraestrutura da Internet no Brasil.

cgi.br

O Comitê Gestor da Internet no Brasil (<https://cgi.br/>), responsável por estabelecer diretrizes estratégicas relacionadas ao uso e desenvolvimento da Internet no Brasil, coordena e integra todas as iniciativas de serviços Internet no País, promovendo a qualidade técnica, a inovação e a disseminação dos serviços ofertados.