JOÃO ESTRELLA

CHECKLIST · SEGURANÇA, GOVERNANÇA E RESPONSABILIDADE

Antes de colocar no ar

Cinco blocos pra rodar antes de publicar qualquer coisa feita com IA: site, automação, agente, planilha compartilhada. Leva três minutos e evita os cinco jeitos mais comuns de se machucar. A ordem dos blocos é a ordem do estrago: começa no que não dá pra desfazer.

COMO USAR

Roda inteira, toda vez. Checklist que você preenche de memória não serve, porque o item que você esquece é justamente o que você não lembrou. Imprima ou deixe aberto ao lado.

1. Segredos

- [ ] **Nenhuma chave de API, senha ou token dentro dos arquivos do projeto**
  Chave de API é a senha longa que identifica a sua conta num serviço. Ela vive numa variável de ambiente, que é uma configuração guardada fora dos arquivos e lida na hora de rodar.
- [ ] **Conferi também o histórico, não só o arquivo atual**
  Segredo apagado depois continua acessível no registro antigo de alterações.
- [ ] **Se algo vazou: revoguei a chave, não só apaguei**
  Existem robôs varrendo repositórios públicos atrás de padrão de chave, e a janela entre subir e ser encontrado se mede em minutos. Revogar invalida. Apagar, não.

2. Dados

- [ ] **Nenhum dado pessoal de cliente ou de terceiro no que vai ficar público**
- [ ] **Quem aparece nominalmente autorizou aparecer**
- [ ] **Nenhum arquivo de teste com dado real esquecido na pasta**
  É o vazamento mais bobo e mais frequente: a planilha real usada pra testar, que sobe junto com o resto.
- [ ] **Se conectei uma base de documentos à IA, abri e olhei o que tem dentro**
  Assistente conectado a uma pasta não entende hierarquia: ele responde qualquer pergunta com o que achou. Confira também quem pode consultar e quem pode adicionar documento, porque contaminar a base contamina todas as respostas.

3. Permissões

- [ ] **O agente tem só o acesso que a tarefa exige**
  Agente é a IA que executa ação, não só devolve texto. O estrago possível é do tamanho da permissão, não do tamanho da tarefa. Leitura antes de escrita. Uma pasta antes do computador inteiro.
- [ ] **Ação irreversível pede aprovação: apagar, enviar pra fora, mexer em dinheiro**
- [ ] **Revisei o que está conectado e desconectei o que não uso mais**
  Acesso concedido tende a ficar aberto pra sempre, porque ninguém lembra de fechar.

4. Custo

- [ ] **Teto de gasto configurado no serviço**
  Cobrança por uso não tem freio natural. Você não acaba o crédito e para: você continua, e a conta continua.
- [ ] **Alerta de consumo ativo, avisando quando passar de um valor**
- [ ] **Todo loop tem limite de repetições, não só condição de sucesso**
  "Tente no máximo 5 vezes e depois pare e me avise." O que uma pessoa faria 50 vezes num dia, um loop faz 50 mil.

5. Conteúdo externo

- [ ] **Sei de onde vem tudo que o agente lê**
  E-mail recebido, página da web, PDF de terceiro, comentário de cliente, currículo enviado. Todo conteúdo externo é potencialmente uma instrução.
- [ ] **Nada que lê conteúdo de estranho executa ação irreversível sozinho**
  É a única defesa prática contra instrução escondida em texto de terceiro, e o setor ainda não tem solução fechada pra isso. Essa é a resposta honesta.

Conferência final

O ÚLTIMO PASSO, E O QUE MAIS GENTE PULA

Abri o resultado como um estranho abriria, e olhei o que dá pra ver.

Sem estar logado, do celular, na aba anônima. É o passo mais barato da lista, e o único que nenhuma ferramenta faz por você.

A frase pra levar

COMO OS RISCOS SE CONECTAM

Instrução escondida em texto de terceiro é a arma. Permissão excessiva é o que transforma a arma em estrago.

Por isso os blocos 3 e 5 andam juntos: se o agente pode ler e-mail de estranho e também pode enviar dinheiro, o problema não é o e-mail, é a combinação.

Fontes: OWASP, "OWASP GenAI LLM Top 10 2026", e a tradução oficial em português da edição 2025, "OWASP Top 10 para Aplicações de LLM e IA Generativa". Os itens deste checklist saem de cinco riscos dessa lista: prompt injection, divulgação de informação sensível, agência excessiva, fraquezas em base de documentos conectada e consumo sem limite. O código de cada um muda de uma edição pra outra, então aqui eles vão pelo nome. Anthropic, "Use Claude Cowork safely", e a página de segurança da documentação do Claude Code. OpenAI, "Safety best practices".