Golpes · Uso policial · LGPD

o Cefet/RJ agora é UFCIRJ

O bem e o mal com LLMs

A mesma tecnologia que escreve um golpe pode ajudar a investigá-lo?

LLM (Large Language Model): modelo de linguagem de grande escala, capaz de gerar, resumir e classificar texto.

1
Enganar
golpes em escala

→

2
Investigar
defesa e relatórios policiais

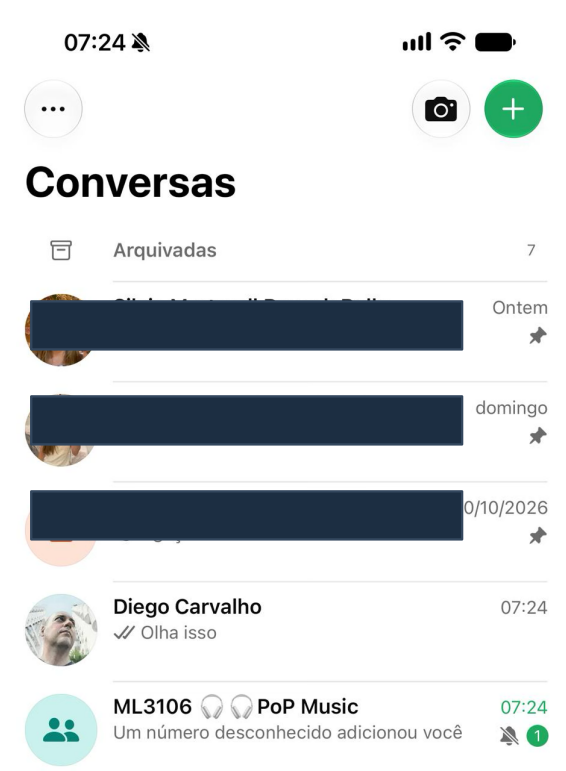
→

3
Proteger
privacidade e LGPD

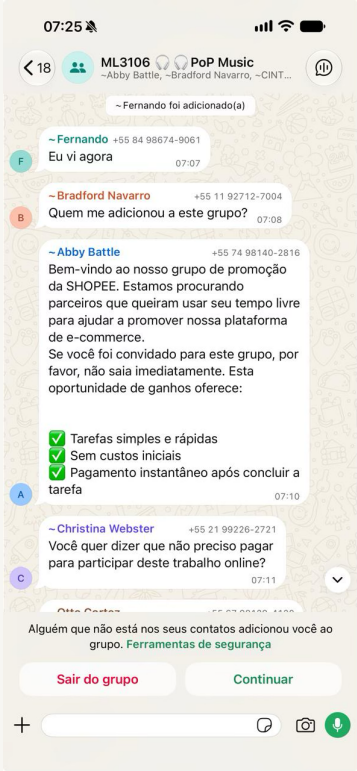
Diego Carvalho · Universidade Federal de Ciência e Inovação do Rio de Janeiro - UFCIRJ

Motivação

Um grupo de WhatsApp para golpes usando um robô com um LLM







WhatsApp

Robô usando IA para golpes em diversos idiomas

Características

- Golpe de pirâmide
- Baixo valor (R$ 8,00)
- Venda de produtos

Fonte: Pedro Henrique Rausch Bello - PPPRO

Produzir texto não é comprovar fatos

O que o modelo consegue fazer não diz nada sobre a finalidade, a veracidade ou a segurança do resultado.

Capacidade explorada	Garantia que não devemos presumir
Personalizar mensagens	Boa intenção de quem usa
Resumir relatos e áudios	Fidelidade integral à fonte
Gerar texto plausível e fluente	Veracidade das afirmações
Reproduzir informações vistas	Proteção automática dos dados pessoais

Capacidade ≠ garantia: é essa distância que golpes exploram e que sistemas públicos precisam controlar.

Fonte: Heiding et al., IEEE Access (2024) · Ferguson, Nw. U. L. Rev. (2025) · Carlini et al., USENIX Security (2021)

O golpe deixou de ser genérico

o Cefet/RJ agora é UFCIRJ

Phishing genérico

EXEMPLO FICTÍCIO

"Sua conta foi bloqueada. Clique aqui para regularizar."

Spear phishing (direcionado)

EXEMPLO FICTÍCIO

"Olá, **estudante**. Há uma pendência na sua inscrição no **evento do curso** de **Ciência da Computação**. Confira seus dados neste endereço."

Personalização
usa contexto real da vítima

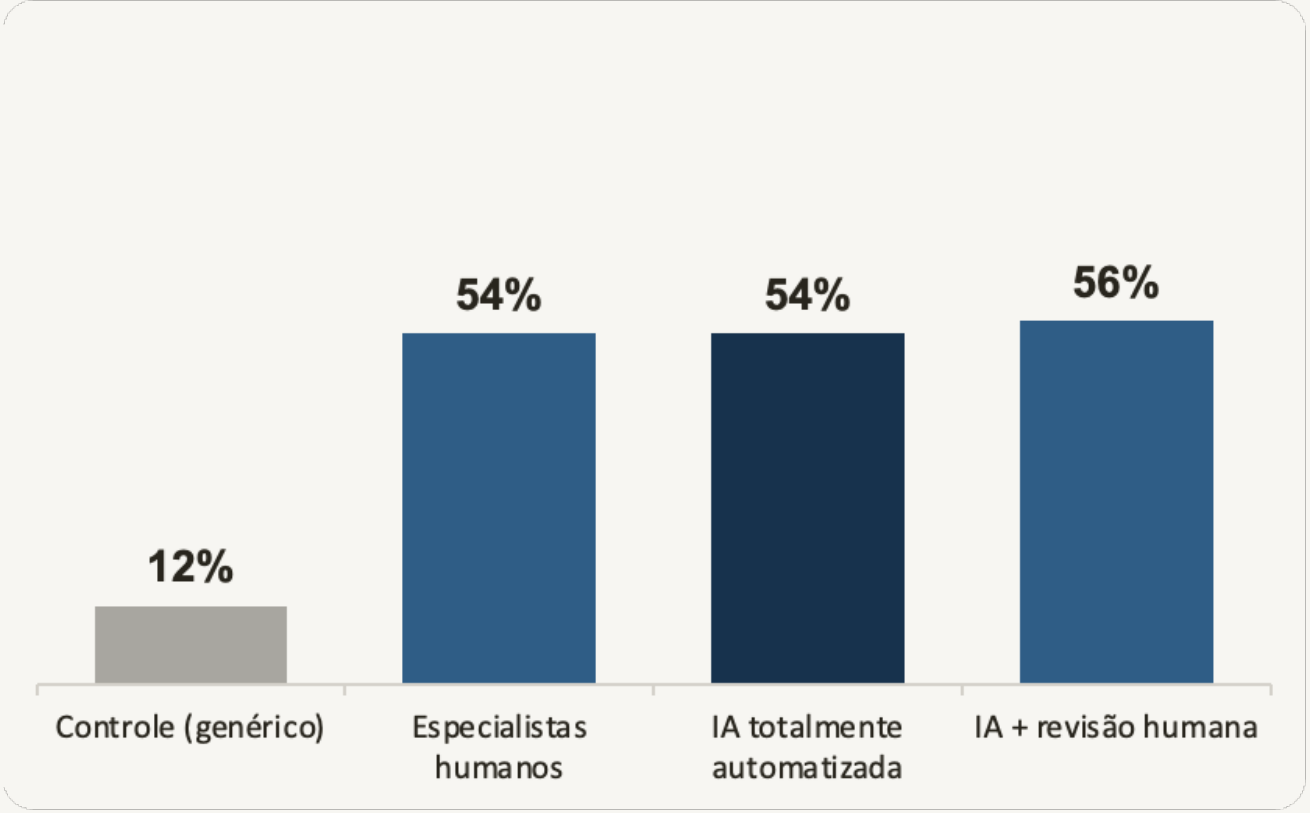
Automação
gera centenas de variações

Menor custo
spear phishing quase ao preço do genérico

Fonte: Heiding, Schneier et al., "Devising and Detecting Phishing Emails Using LLMs", IEEE Access (2024)

Automação total já empata com especialistas

Taxa de clique em e-mails de spear phishing, por grupo (101 participantes)



4,5×

mais cliques da IA automatizada (54%) do que no controle (12%)

Como ler o resultado

- Um experimento específico; preprint
- Clique ≠ senha entregue ou prejuízo
- 54% vs. 56% não indica superioridade

Fonte: Heiding, Lermen, Kao, Schneier e Vishwanath, arXiv:2412.00586 (2024, preprint)

A mesma IA também desarma o golpe

o Cefet/RJ agora é UFCIRJ

Mensagem analisada

EXEMPLO FICTÍCIO

“Há uma pendência na sua inscrição no evento do curso. Confira seus dados neste endereço.”

Sinal 1 · cria urgência e pede uma ação

Sinal 2 · invoca um contexto acadêmico real

Sinal 3 · leva a um endereço externo

Mensagem suspeita
chega ao usuário

↓

Análise assistida
o LLM aponta sinais e explica

↓

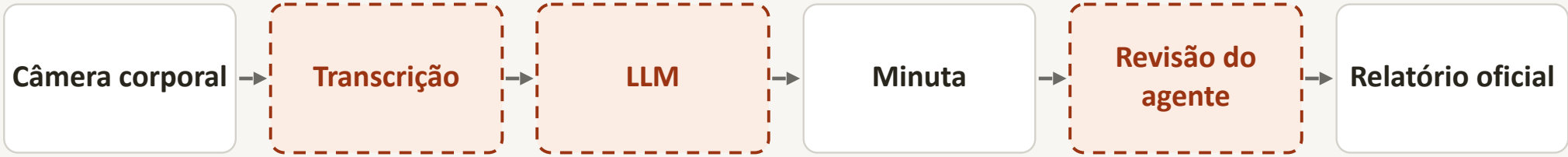
Verificação independente
canal oficial confirma ou não

Desempenho varia muito. No estudo da IEEE Access (2024), o melhor modelo reconheceu a intenção maliciosa em só 25% dos e-mails gerados por GPT; no preprint seguinte, o Claude 3.5 Sonnet passou de 90%, com poucos falsos positivos. Estudos diferentes, não comparáveis diretamente: **“parece seguro” não é certificado.**

Fonte: Heiding et al., IEEE Access (2024) · Heiding et al., arXiv:2412.00586 (2024)

Da câmera corporal à narrativa oficial

o Cefet/RJ agora é UFCIRJ



Contorno tracejado = ponto em que o relato pode ser transformado: erro de transcrição, acréscimo do modelo ou revisão superficial.

Benefício pretendido

Menos tempo gasto com a redação de relatórios.

Risco central

Erros, vieses e lacunas preenchidas pelo modelo viram a narrativa oficial — que orienta denúncia, prisão preventiva e acordos.

Pergunta-chave: cada afirmação pode ser conferida na gravação?

Fonte: Ferguson, "Generative Suspicion and the Risks of AI-Assisted Police Reports", Northwestern University Law Review (2025)

Texto convincente não é evidência

EXEMPLO FICTÍCIO

O que diz a fonte (áudio)

"Eu **não consegui identificar** o objeto na mão dele."

→

O que diz o resumo gerado

"O indivíduo **portava uma arma**."

Incerteza removida → **Afirmação acrescentada** → **Significado jurídico alterado**

Requisito de projeto: rastrear cada frase do relatório até o trecho da gravação que a sustenta.

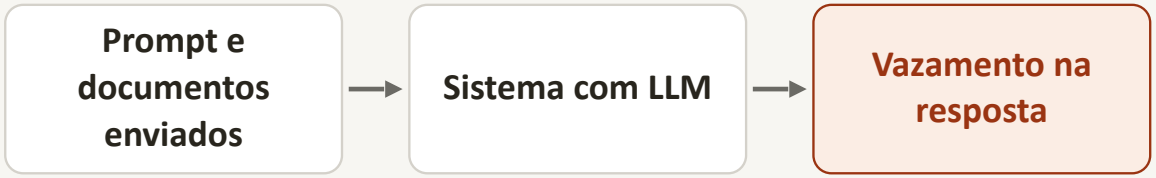
Fonte: Exemplo ilustrativo; riscos de alucinação e preenchimento de lacunas discutidos em Ferguson, Nw. U. L. Rev. (2025)

Estar no modelo não é estar anonimizado

Caminho 1 · Treinamento



Caminho 2 · Aplicação



São riscos distintos: nem toda conversa altera os parâmetros do modelo.

Carlini et al. · USENIX Security 2021

Extraíram do GPT-2 centenas de trechos memorizados do treinamento, incluindo nomes, telefones e e-mails.

LLM-PBE · PVLDB 2024

Avalia privacidade no ciclo inteiro: pré-treinamento, ajuste fino e prompts, com ataques e defesas.

Demonstra o mecanismo, não uma taxa de vazamento dos chatbots atuais.

Fonte: Carlini et al., USENIX Security (2021) · Li et al., LLM-PBE, PVLDB 17(11) (2024)

Segurança pública não é cheque em branco

LGPD — Lei nº 13.709/2018 (Lei Geral de Proteção de Dados Pessoais)

Aplicações sujeitas à LGPD

ex.: chatbot de banco, sistema acadêmico

- Finalidade e base legal (consentimento é só uma das bases)
- Necessidade: tratar só o indispensável
- Segurança, transparência e prestação de contas
- Art. 20: revisão de decisões unicamente automatizadas

Fins exclusivos de segurança pública ou investigação penal

art. 4º, III — exceção à aplicação da LGPD

- Regidos por legislação específica (§ 1º)
- Medidas proporcionais e estritamente necessárias
- Devido processo legal
- Princípios gerais de proteção e direitos do titular

O regime depende da finalidade do tratamento, não da instituição.

Fonte: Lei nº 13.709/2018, arts. 4º, 6º e 20 (Planalto)

Antes de pôr um LLM em produção

o Cefet/RJ agora é UFCIRJ

1 · Dados

Quais informações entram e por que são necessárias?

2 · Verificação

Como conferir cada saída na fonte original?

3 · Responsabilidade

Quem responde pelo uso e pelo erro?

4 · Contestação

Como a pessoa afetada corrige ou questiona o resultado?

Gerar texto não é comprovar fatos — nem assumir responsabilidade.

Fonte: Ferguson (2025) · LLM-PBE (2024) · LGPD (2018)

Obrigado

o Cefet/RJ agora é UFCIRJ

1 · Contato

diego.carvalho@cefet-rj.br

Aviso legal

LLMs foram utilizados extensamente na preparação desses slides.

3 · Ressalva

Foi utilizado o www.perplexity.ai para revisão gramatical, brainstorming, diagramação dos slides, etc.

4 · Responsabilidade humana

O conteúdo foi revisado, validado e assumo total responsabilidade pela exatidão, originalidade e ética do conteúdo.

2026 (C) Diego Carvalho
Universidade Federal de Ciência e Inovação do Rio de Janeiro