Riscos de Privacidade e Segurança da Informação em Sistemas de Inteligência Artificial e IA Generativa: Um Estudo Empírico sobre Percepções de Profissionais e Práticas de Mitigação no Brasil

Luciane de Andrade Oliveira Sales
Universidade de Brasília (UnB)
Brasília, Distrito Federal, Brasil
lucianesales@gmail.com

Edna Dias Canedo
Universidade de Brasília (UnB)
Brasília, Distrito Federal, Brasil
ednacanedo@unb.br

RESUMO

Contexto: A crescente adoção de Inteligência Artificial (IA) e Inteligência Artificial Generativa (GenAI) nas organizações tem intensificado preocupações relacionadas à privacidade e à segurança da informação, que vão além de vulnerabilidades técnicas e abrangem fatores sociotécnicos, organizacionais e regulatórios. **Objetivo**: Este estudo investiga os riscos de privacidade e segurança da informação associados a sistemas de IA e GenAI, examinando as percepções de profissionais, as estratégias de mitigação adotadas e os desafios que limitam sua efetividade na prática. **Método**: Foi conduzido um estudo de métodos mistos combinando uma revisão da literatura com um survey aplicado a 101 profissionais do setor público brasileiro. Os dados quantitativos foram analisados por meio de estatísticas descritivas, enquanto as respostas qualitativas foram examinadas utilizando codificação indutiva. **Resultados**: Os profissionais percebem os riscos de privacidade e segurança da informação como altamente críticos, destacando-se especialmente o vazamento de dados, a não conformidade legal, a falta de transparência, ataques de *prompt injection* e o uso malicioso de conteúdos gerados por IA. Embora estratégias de mitigação sejam amplamente adotadas, sua efetividade é percebida como limitada devido a lacunas de governança, baixa maturidade organizacional, treinamento insuficiente, uso de *shadow IA* e transparência limitada por parte dos fornecedores. **Conclusões**: Os resultados revelam um desalinhamento entre a adoção de estratégias de mitigação e a confiança em sua efetividade, reforçando que os riscos de privacidade e segurança da informação em sistemas de IA são fundamentalmente sociotécnicos e requerem abordagens integradas que combinem soluções técnicas, organizacionais e de governança.

KEYWORDS

Inteligência Artificial, IA Generativa, Riscos de Privacidade, Riscos de Segurança da Informação, Estudo Empírico, Percepção de Profissionais, Mitigação de Riscos

1 INTRODUÇÃO

Sistemas de Inteligência Artificial (IA) tornaram-se parte integrante de uma ampla gama de aplicações organizacionais, governamentais e sociais, baseando-se extensivamente em grandes volumes de dados e em processos automatizados de tomada de decisão. Como consequência, preocupações relacionadas à privacidade e à segurança da informação emergiram como desafios centrais para a adoção confiável de tecnologias de IA [2, 6, 12]. Pesquisas anteriores demonstram consistentemente que os riscos de privacidade e segurança da informação em sistemas de IA não se limitam ao acesso não autorizado a dados, mas incluem também identificação indireta, inferência de atributos sensíveis, falta de transparência e não conformidade regulatória ao longo do ciclo de vida da IA [2, 6, 11, 18, 29].

A recente adoção em larga escala da Inteligência Artificial Generativa (GenAI), especialmente dos Modelos de Linguagem de Grande Escala (LLMs), intensificou ainda mais essas preocupações [19]. Diferentemente dos sistemas tradicionais de IA, aplicações baseadas em LLMs expõem capacidades avançadas de inferência por meio de interfaces interativas em linguagem natural, aumentando substancialmente a superfície de ataque para violações de privacidade e segurança da informação [21, 27]. Estudos empíricos e baseados em surveys demonstram que LLMs são vulneráveis a riscos como inferência de pertencimento (*membership inference*), extração de dados de treinamento, inversão de modelos, vazamento contextual de dados e ataques baseados em *prompts*, mesmo em cenários de caixa-preta [7, 20]. Esses riscos desafiam pressupostos estabelecidos sobre minimização de dados, uso controlado de informações e separação clara entre ameaças no treinamento e na fase de inferência [1].

Além das vulnerabilidades técnicas, fatores organizacionais e humanos desempenham papel significativo na forma como riscos de privacidade e segurança da informação são percebidos e gerenciados [15]. Evidências empíricas provenientes de equipes de desenvolvimento de software indicam que profissionais geralmente estão cientes dos riscos de privacidade e segurança da informação associados ao uso de LLMs; contudo, as estratégias de mitigação frequentemente se baseiam em práticas informais, como evitar entradas sensíveis ou aplicar anonimização ad hoc, em vez de controles organizacionais sistemáticos [9]. De maneira semelhante, estudos com tomadores de decisão em nível executivo mostram que preocupações relacionadas à privacidade, segurança da informação e conformidade regulatória permanecem como importantes barreiras à confiança e à adoção de sistemas de IA em contextos organizacionais [14].

Em resposta a esses desafios, estruturas regulatórias e de governança têm enfatizado princípios como *privacy by design*, responsabilidade, transparência e gestão de riscos [5, 25, 32]. Por exemplo, o *NIST AI Risk Management Framework* [26] fornece orientações estruturadas para identificação e gestão de riscos sistêmicos associados a sistemas de IA. Entretanto, os frameworks existentes são em grande parte agnósticos à tecnologia e oferecem orientações operacionais limitadas adaptadas às características específicas de sistemas de GenAI, especialmente aquelas relacionadas à inferência interativa, uso secundário de dados e vulnerabilidades baseadas em *prompts*.

Embora trabalhos anteriores forneçam taxonomias relevantes de riscos de privacidade e segurança da informação em IA e GenAI [2, 13, 18, 22, 23, 29–31], as evidências permanecem fragmentadas entre perspectivas técnicas, organizacionais e regulatórias, com limitada conexão empírica com as decisões cotidianas de mitigação adotadas por profissionais. Para abordar essa lacuna, investigamos os riscos de privacidade e segurança da informação em sistemas de IA e GenAI por meio de (i) uma revisão da literatura e (ii) um survey com 101 profissionais que atuam com privacidade e segurança da informação, predominantemente do setor público brasileiro. Nossas contribuições incluem: uma visão integrada de riscos e estratégias de mitigação que abrangem dimensões técnicas e de governança; evidências empíricas sobre a criticidade percebida pelos profissionais, a adoção e a eficácia percebida das medidas de mitigação; e insights qualitativos que explicam por que medidas amplamente adotadas ainda são percebidas como insuficientes na prática.

2 CONTEXTO E TRABALHOS RELACIONADOS

Privacidade e Proteção de Dados em Sistemas de IA. Os riscos de privacidade são amplamente reconhecidos como inerentes aos sistemas de IA devido à sua natureza orientada a dados e à sua capacidade de inferir informações sensíveis além dos dados explicitamente fornecidos. Trabalhos anteriores adotam uma perspectiva de ciclo de vida para demonstrar que ameaças à privacidade emergem durante a coleta de dados, o treinamento de modelos, a implantação e a operação dos sistemas. Um survey abrangente de [29] classifica esses riscos em identificação, tomada de decisão imprecisa, falta de transparência e não conformidade regulatória, enfatizando que ameaças à privacidade decorrem não apenas de ataques técnicos, mas também de práticas organizacionais e falhas de governança [11].

No contexto de GenAI e Modelos de Linguagem de Grande Escala, surveys relatam maior exposição a ataques de inferência de pertencimento, inversão de modelos, extração de dados de treinamento e vazamento contextual de informações [20, 22]. Estudos mais recentes consolidam esses riscos, destacando efeitos de memorização, exposição de dados secundários e limitações persistentes de transparência como desafios centrais em sistemas de GenAI [13]. Embora estratégias de mitigação como privacidade diferencial, aprendizado federado e implantação segura sejam frequentemente discutidas, elas geralmente envolvem trade-offs entre privacidade, utilidade e desempenho, além de serem difíceis de operacionalizar em larga escala.

Evidências empíricas indicam que riscos de privacidade na prática frequentemente são gerenciados por meio de comportamentos informais, em vez de controles organizacionais sistemáticos [3]. Por exemplo, um survey com equipes brasileiras de desenvolvimento de software mostra que profissionais estão principalmente preocupados com vazamento de dados e uso indevido de dados pessoais, enquanto a mitigação frequentemente se baseia em evitar entradas sensíveis e aplicar anonimização ad hoc [9]. Padrões semelhantes são observados em estudos centrados em desenvolvedores. Kapitsaki et al. [16] demonstram que a conformidade com privacidade frequentemente é tratada de forma reativa e fragmentada, refletindo lacunas no entendimento da legislação e limitada integração de requisitos de privacidade nas práticas de desenvolvimento. De modo geral, a literatura converge para a necessidade de tratar a privacidade como uma propriedade sistêmica dos sistemas de IA, mas permanece predominantemente técnica e oferece insights limitados sobre como esses riscos são operacionalizados em contextos organizacionais reais, especialmente em aplicações de GenAI já implantadas.

Riscos de Segurança da Informação em IA e GenAI. Os riscos de segurança da informação em sistemas de IA estão intimamente relacionados às preocupações com privacidade, uma vez que muitos ataques comprometem simultaneamente confidencialidade, integridade e disponibilidade. Surveys clássicos documentam ameaças como envenenamento de dados, manipulação adversarial, evasão e extração de modelos, destacando como vulnerabilidades em algoritmos de aprendizado podem resultar em consequências prejudiciais e exposição de dados sensíveis. Pesquisas recentes mostram que GenAI introduz riscos qualitativamente novos, particularmente por meio da interação baseada em *prompts*, que emergiu como uma superfície crítica de ataque [7]. Ameaças como *prompt injection*, ataques indiretos por *prompt*, sequestro de objetivos (*goal hijacking*) e divulgação não autorizada exploram o raciocínio contextual e a interação em múltiplos turnos em interfaces de linguagem natural.

Surveys em larga escala indicam ainda que riscos de segurança da informação em sistemas de GenAI são sistêmicos e interdependentes, abrangendo ataques em tempo de inferência, uso indevido de conteúdo gerado e falhas de governança que vão além de modelos tradicionais de ameaça puramente técnicos [1, 6]. Pesquisas complementares sobre segurança de LLMs reforçam que vulnerabilidades na fase de inferência tornaram-se tão críticas quanto ataques durante o treinamento e enfatizam que muitas técnicas de mitigação ainda são experimentais, carecem de padronização ou não escalam adequadamente para implantações reais [20]. Em geral, a literatura destaca que riscos de segurança da informação são frequentemente tratados isoladamente de controles organizacionais, deixando lacunas persistentes em responsabilidade e gestão operacional de riscos.

Cenário Regulatório e Ético. Além das vulnerabilidades técnicas, sistemas de IA e GenAI operam em ambientes regulatórios e éticos cada vez mais complexos. Regulamentações de proteção de dados e frameworks emergentes de governança de IA enfatizam princípios como *privacy by design*, transparência, responsabilidade e supervisão humana. Contudo, traduzir esses requisitos de alto nível em práticas concretas de engenharia e operação continua sendo um desafio. Estudos orientados à governança e centrados no ser humano argumentam que riscos de privacidade e segurança da informação não podem ser mitigados de forma eficaz sem considerar dimensões sociais e organizacionais, alertando que dependência excessiva de salvaguardas técnicas pode obscurecer responsabilidades e comprometer a confiança [20]. Evidências empíricas de tomadores de decisão organizacionais mostram ainda que preocupações relacionadas à privacidade, segurança da informação e conformidade constituem barreiras centrais à adoção de IA, especialmente no setor público [14].

Frameworks de gestão de riscos, incluindo modelos de maturidade e abordagens de governança de IA, evidenciam uma lacuna persistente entre expectativas regulatórias e práticas atuais. Embora estruturas como o *NIST AI Risk Management Framework* forneçam

orientações estruturadas [26], elas permanecem amplamente agnósticas às particularidades da GenAI, incluindo inferência interativa, reutilização de dados e exposição secundária de informações. Trabalhos anteriores oferecem contribuições relevantes sobre princípios regulatórios e riscos éticos, mas permanecem fragmentados entre perspectivas técnicas, organizacionais e legais, oferecendo compreensão empírica limitada sobre como tais requisitos são operacionalizados no uso cotidiano de sistemas de IA e GenAI nas organizações. A Tabela 1 consolida esses achados ao sintetizar como estudos anteriores abordam riscos de privacidade e segurança da informação, estratégias de mitigação propostas e suas limitações relatadas em diferentes perspectivas.

3 CONFIGURAÇÃO DO ESTUDO

Este estudo adota um desenho de pesquisa de métodos mistos, combinando uma revisão da literatura com um survey empírico [17]. O componente de survey foi concebido para complementar os achados da literatura, capturando as percepções dos profissionais, as práticas de mitigação adotadas e os desafios relacionados aos riscos de privacidade e segurança da informação no uso de sistemas de Inteligência Artificial (IA) e IA Generativa (GenAI). Tanto a revisão da literatura quanto o survey foram orientados pelas seguintes questões de pesquisa (RQ):

RQ1: Quais riscos de privacidade e segurança da informação estão associados ao uso de sistemas de IA e GenAI?
RQ2: Como esses riscos são percebidos por profissionais e organizações?
RQ3: Quais estratégias de mitigação são reportadas na literatura e na prática?

A RQ1 é abordada por meio da revisão da literatura, que consolida e categoriza os riscos de privacidade e segurança da informação reportados em estudos anteriores, enquanto o survey oferece validação empírica ao captar a avaliação dos profissionais quanto à relevância e criticidade desses riscos. A RQ2 é examinada por meio do survey com base em medidas quantitativas de importância percebida, complementadas por evidências qualitativas que contextualizam tais percepções em cenários organizacionais reais. A RQ3 é investigada por meio de uma análise integrada das estratégias de mitigação identificadas na literatura e daquelas reportadas pelos profissionais, com foco em sua adoção, efetividade percebida e desafios de implementação.

A Figura 1 apresenta uma visão geral do desenho de pesquisa adotado neste estudo. Ela mostra como a revisão da literatura fundamenta o planejamento do survey, incluindo a realização de um teste piloto, e como as análises quantitativa e qualitativa são combinadas para responder às questões de pesquisa. Essa visão geral também explicita a separação entre as evidências derivadas de estudos prévios e aquelas coletadas diretamente com os profissionais, o que sustenta a organização dos resultados em achados baseados na literatura e achados baseados no survey.

Desenho do Survey. O instrumento de survey foi elaborado com base nos riscos, estratégias de mitigação e lacunas identificados na literatura sobre privacidade e segurança da informação em IA e GenAI (Tabela 1). O questionário é composto por quatro partes principais: (i) perfil dos respondentes, (ii) percepção sobre riscos de privacidade e segurança da informação, (iii) estratégias de mitigação e sua efetividade percebida, e (iv) desafios e lacunas na gestão desses riscos. Além disso, foram incluídas duas questões abertas para permitir que os participantes descrevessem situações concretas, preocupações ou desafios relacionados à privacidade e à segurança da informação em seu contexto profissional.

A maior parte das questões foi formulada com escalas Likert de cinco pontos, seguindo práticas comuns em estudos empíricos com diversos profissionais. Os riscos percebidos e os desafios foram medidos por meio de uma escala de importância variando de "Nada importante" a "Muito importante", enquanto as estratégias de mitigação e sua efetividade foram avaliadas por meio de uma escala de concordância variando de "Discordo totalmente" a "Concordo totalmente". O instrumento completo do survey e seu mapeamento para as questões de pesquisa são apresentados na Tabela 2. Antes da coleta principal, foi realizado um estudo piloto com cinco profissionais para avaliar a clareza, a consistência e a extensão do questionário. O feedback recebido levou a pequenos ajustes de redação, com o objetivo de melhorar a compreensão e reduzir ambiguidades, sem alterar a estrutura ou a intenção original das perguntas.

Coleta de Dados. A população-alvo do survey é composta por profissionais que atuam com privacidade e segurança da informação no Brasil e que utilizam, gerenciam ou estão expostos a sistemas de IA e GenAI em suas atividades profissionais. Os participantes foram recrutados por meio das redes de contato profissionais e acadêmicas dos autores, bem como por meio de publicações compartilhadas em redes sociais profissionais, seguindo uma estratégia de amostragem por conveniência, comumente adotada em estudos empíricos exploratórios. O survey foi aplicado online por meio da plataforma Google Forms e permaneceu disponível para respostas por um período de 45 dias. A participação foi voluntária, e nenhuma informação pessoal identificável foi coletada. O tempo médio de preenchimento do questionário foi de aproximadamente 12 minutos.

Antes de acessar o questionário, os participantes foram apresentados a um termo de consentimento livre e esclarecido, descrevendo os objetivos do estudo, o caráter voluntário da participação e as medidas adotadas para garantir o anonimato e a conformidade com as regulamentações aplicáveis de proteção de dados, incluindo a Lei Geral de Proteção de Dados Pessoais (LGPD) [5].

Análise dos Dados. Os dados quantitativos foram analisados por meio de técnicas de estatística descritiva. Para as questões em escala Likert, foram calculadas distribuições de frequência, percentuais e medidas de tendência central, com o objetivo de sintetizar as percepções dos participantes sobre riscos de privacidade e segurança da informação, estratégias de mitigação e desafios enfrentados. Quando apropriado, os resultados foram estratificados de acordo com o perfil dos respondentes, como área de atuação e experiência com sistemas de IA ou GenAI, a fim de identificar possíveis diferenças entre grupos.

As respostas às questões abertas foram analisadas qualitativamente por meio de uma abordagem de codificação indutiva [4]. A análise concentrou-se na identificação de temas recorrentes relacionados a preocupações com privacidade e segurança da informação, práticas de mitigação e limitações percebidas. Os achados qualitativos foram utilizados para complementar e contextualizar os resultados quantitativos, favorecendo uma interpretação mais abrangente dos padrões observados.

Tabela 1: Síntese de trabalhos relacionados sobre riscos de privacidade e segurança da informação em sistemas de IA e IA Generativa

Estudo	Foco	Riscos Identificados	Estratégias de Mitigação	Principais Limitações
[29]	Ciclo de vida da IA, governança de privacidade	Riscos de identificação, decisões imprecisas, falta de transparência, não conformidade regulatória	PETs (Tecnologias de Aprimoramento de Privacidade), privacy by design, controles de processos e políticas	Cobertura limitada de riscos específicos de GenAI/LLMs e de ataques baseados em prompts
[20]	Privacidade em LLMs	Inferência de pertencimento (membership inference), inversão de modelo, extração de dados de treinamento, vazamento contextual	Privacidade diferencial (DP), treinamento seguro, implantação controlada	Predominantemente técnico; limitada operacionalização organizacional e regulatória
[7]	Interação baseada em prompts com LLMs	Injeção de prompts, ataques indiretos por prompt, sequestro de objetivos (goal hijacking), divulgação não autorizada	Defesas contra prompt injection, controle de acesso, salvaguardas em tempo de inferência	Evidências limitadas sobre eficácia em larga escala de implantação
[9]	Percepções de profissionais (LLMs)	Vazamento de dados confidenciais, uso indevido de dados pessoais, riscos de terceiros, lacunas de transparência	Evitar dados sensíveis, anonimização, uso de ferramentas confiáveis e políticas organizacionais	Práticas autorrelatadas; limitada profundidade organizacional
[14]	Percepções de executivos (C-level) e barreiras à adoção	Privacidade, segurança, conformidade e responsabilização como inibidores de confiança	Governança, treinamento e controles organizacionais (nível estratégico)	Não específico para GenAI; detalhes técnicos limitados
[26]	Governança de riscos em IA	Riscos sistêmicos de privacidade, segurança e confiabilidade	Processos de gestão de riscos, estruturas de governança, monitoramento	Abordagem agnóstica à tecnologia; orientação operacional limitada para GenAI
[16]	Discussões de desenvolvedores sobre regulação de privacidade	Interpretação equivocada de obrigações legais, incerteza sobre conformidade, lacunas entre regulação e prática	Compartilhamento informal de conhecimento, discussões entre pares, interpretação comunitária	Evidência limitada a discussões online; ausência de validação organizacional
[10]	IA em tomada de decisão estratégica e organizacional	Perda de responsabilização, opacidade em decisões baseadas em IA, dependência excessiva da automação	Supervisão humana, mecanismos de governança, alinhamento estratégico	Perspectiva de alto nível; foco limitado em riscos técnicos
[24]	Desafios de privacidade e segurança em sistemas de IA	Vazamento de dados, riscos de inferência, falta de transparência, vulnerabilidades de segurança	Salvaguardas técnicas, princípios de privacy by design	Evidências empíricas limitadas provenientes de profissionais
[8]	Governança de IA centrada no ser humano	Uso indevido ético, erosão da confiança, supervisão humana insuficiente	Governança centrada no ser humano, estruturas éticas e participativas	Foco conceitual; orientação operacional limitada
[28]	Análise empírica de riscos de IA na prática	Violações de privacidade, uso indevido de segurança, lacunas de governança	Políticas, treinamento e controles organizacionais	Foco limitado em GenAI e riscos baseados em prompts
[22]	Survey sobre privacidade em modelos GenAI	Vazamento de dados, memorização, ataques de inferência, falta de transparência	Aprendizado com preservação de privacidade, técnicas de mitigação de riscos	Baseado em survey; carece de evidência organizacional empírica
[13]	Privacidade e segurança em IA Generativa	Vazamento de dados, abuso de prompts, uso indevido, ameaças de conformidade e segurança	Salvaguardas técnicas, governança e medidas de políticas	Escopo amplo; limitada profundidade sobre práticas organizacionais
[6]	Segurança e privacidade em LLMs	Ataques de inferência, vazamento de dados, uso indevido, ameaças adversariais	Security-by-design, técnicas de mitigação de riscos	Foco em survey; ausência de validação com profissionais
[1]	Ameaças e práticas responsáveis em LLMs	Injeção de prompts, uso indevido, vazamento de dados, falhas de governança	Práticas de IA responsável, salvaguardas e mecanismos de governança	Conceitual; validação empírica limitada

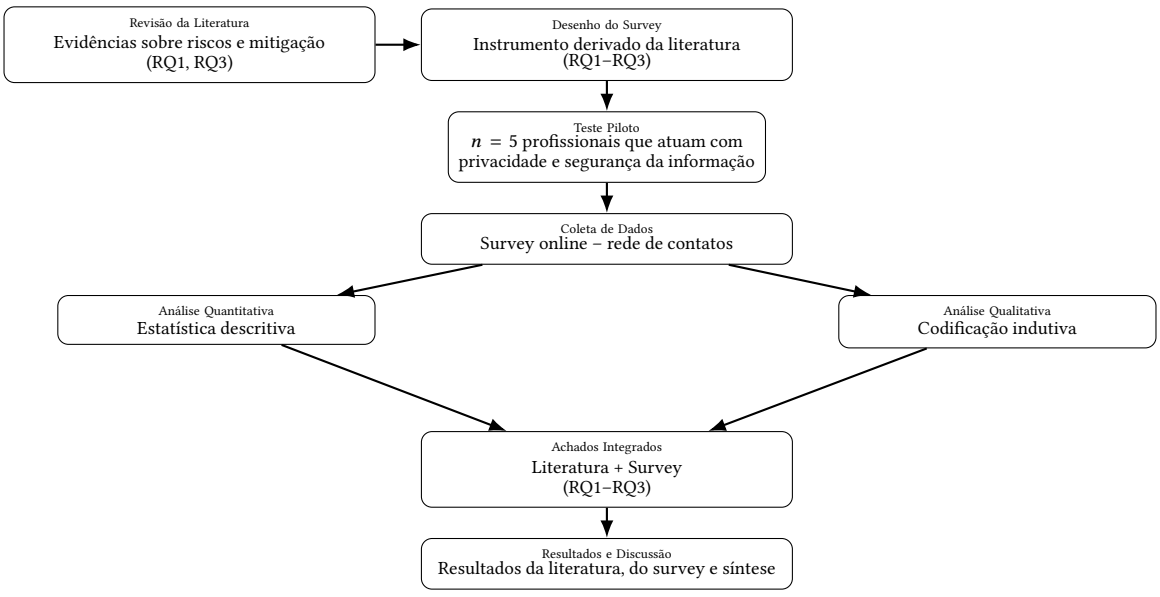

Figura 1: Visão geral do método de pesquisa e seu alinhamento com as questões de pesquisa.

Tabela 2: Questões do survey e mapeamento para as questões de pesquisa

ID	Questão do Survey	RQ
Q1	Em que tipo de organização ou entidade pública você trabalha atualmente?	–
Q2	Qual é o seu nível mais alto de escolaridade?	–
Q3	Qual é a sua área de atuação profissional dentro da organização ou entidade?	–
Q4	Quantos anos de experiência você possui com Inteligência Artificial (IA) ou sistemas orientados a dados?	–
Q5	Você atualmente utiliza ou gerencia ferramentas de GenAI (por exemplo, Large Language Models) em suas atividades profissionais?	–
Q6	Indique o nível de importância que você atribui aos seguintes **riscos de privacidade** associados ao uso de sistemas de IA e GenAI: vazamento de dados pessoais ou sensíveis; inferência de atributos privados a partir de dados não sensíveis; memorização de dados de treinamento por modelos de IA; falta de transparência sobre o uso de dados; não conformidade com regulações de proteção de dados (por exemplo, LGPD).	RQ2
Q7	Indique o nível de importância que você atribui aos seguintes **riscos de segurança da informação** associados ao uso de sistemas de IA e GenAI: envenenamento de dados; extração ou inversão de modelos; *prompt injection* ou ataques indiretos por *prompt*; acesso não autorizado a saídas geradas por IA; uso indevido ou malicioso de conteúdo gerado por IA.	RQ2
Q8	Indique seu nível de concordância com as seguintes afirmações sobre a **adoção de estratégias de mitigação** para riscos de privacidade e segurança da informação no uso de sistemas de IA e GenAI: evitar o uso de dados sensíveis; uso de técnicas de anonimização ou pseudonimização; existência de políticas ou diretrizes organizacionais; oferta de treinamento sobre privacidade, segurança da informação e uso responsável de IA; adoção de controles técnicos de segurança da informação e privacidade.	RQ3
Q9	Indique seu nível de concordância com as seguintes afirmações sobre a **efetividade das estratégias atuais de mitigação**: as estratégias adotadas são suficientes para mitigar riscos de privacidade; as estratégias adotadas são suficientes para mitigar riscos de segurança da informação no uso de sistemas de IA e GenAI.	RQ3
Q10	Indique o nível de importância dos seguintes **desafios e lacunas** na gestão de riscos de privacidade e segurança da informação relacionados à IA e GenAI: falta de conhecimento técnico ou treinamento; ausência de regulações ou diretrizes claras; dificuldade em equilibrar privacidade e segurança da informação com desempenho e inovação; baixa maturidade organizacional ou ausência de estruturas de governança; falta de transparência por parte dos fornecedores de soluções de IA.	RQ2
Q11	Descreva, se aplicável, uma situação concreta, preocupação ou desafio relacionado a **riscos de privacidade** no uso de IA ou GenAI em seu contexto profissional.	RQ2
Q12	Descreva, se aplicável, uma situação concreta, preocupação ou desafio relacionado a **riscos de segurança da informação** no uso de IA ou GenAI em seu contexto profissional.	RQ2

Esses procedimentos analíticos permitiram examinar de forma sistemática como os riscos de privacidade e segurança da informação associados a sistemas de IA e GenAI são percebidos e gerenciados na prática, respondendo diretamente às questões de pesquisa deste estudo. A codificação qualitativa foi conduzida pelos autores por meio de leituras iterativas e refinamento progressivo das categorias analíticas. A análise priorizou padrões recorrentes e saturação temática, em vez de contagens de frequência.

4 RESULTADOS

Esta seção apresenta os resultados derivados da análise dos estudos selecionados na revisão da literatura e no survey. Os achados estão organizados de acordo com as questões de pesquisa e concentram-se em: (i) os riscos de privacidade e segurança da informação associados a sistemas de IA e GenAI (RQ1) e (ii) as estratégias de mitigação reportadas em estudos anteriores (RQ3). As questões relacionadas às percepções dos profissionais (RQ2) são tratadas exclusivamente por meio dos resultados do survey.

4.1 RQ1. Riscos de Privacidade e Segurança da Informação

A literatura analisada indica que os riscos de privacidade são intrínsecos aos sistemas de IA devido à sua dependência de processamento de dados em larga escala e de capacidades avançadas de inferência. Em sistemas de IA tradicionais e em sistemas GenAI, os riscos de privacidade mais frequentemente reportados incluem vazamento de dados pessoais ou sensíveis, inferência de atributos privados a partir de entradas não sensíveis, memorização de dados de treinamento e falta de transparência quanto ao uso e à origem dos dados [20, 29].

Estudos voltados a Large Language Models destacam que esses riscos são agravados pela capacidade dos modelos de memorizar e reproduzir fragmentos de dados de treinamento, mesmo em cenários de caixa-preta. Evidências empíricas mostram que ataques de inferência de pertencimento (*membership inference*) e extração de dados de treinamento podem ocorrer durante a inferência, desafiando a suposição de que os riscos de privacidade se limitam à fase de treinamento [20]. Além disso, interações contextuais e de múltiplos turnos têm se mostrado capazes de ampliar o vazamento de privacidade, especialmente quando informações sensíveis estão implicitamente codificadas em *prompts* ou no histórico conversacional.

Outra preocupação recorrente relacionada à privacidade na literatura é a não conformidade regulatória. Diversos estudos enfatizam que governança de dados inadequada, documentação insuficiente sobre as fontes de dados e transparência limitada dificultam a demonstração de conformidade com regulações de proteção de dados, como GDPR e LGPD [29]. Esses achados sugerem que os riscos de privacidade não são apenas técnicos, mas também organizacionais e procedimentais. Em paralelo às preocupações com privacidade, a literatura identifica um amplo espectro de riscos de segurança da informação que afetam sistemas de IA e GenAI. Ameaças clássicas, como envenenamento de dados, manipulação adversarial e extração de modelos, continuam presentes em diferentes paradigmas de IA. Esses ataques podem comprometer a integridade do modelo, degradar seu desempenho e expor indiretamente informações sensíveis [27].

Estudos recentes demonstram que sistemas GenAI introduzem novas vulnerabilidades de segurança da informação intimamente relacionadas à sua natureza interativa. Em particular, a interação baseada em *prompts* emergiu como uma superfície crítica de ataque. As taxonomias propostas para ameaças baseadas em *prompts* identificam riscos como *prompt injection*, ataques indiretos por *prompt*, sequestro de objetivo (*goal hijacking*) e divulgação não autorizada de informações por meio de saídas geradas [7]. Diferentemente dos ataques tradicionais, essas ameaças muitas vezes não exigem acesso aos dados de treinamento ou aos parâmetros do modelo, o

que dificulta sua detecção e prevenção com controles convencionais de segurança da informação. A literatura também destaca que os riscos de privacidade e segurança da informação estão profundamente interconectados, como mostrado na Tabela 1. Violações de segurança, como acesso não autorizado às saídas do modelo ou aos pontos de inferência, podem levar diretamente a violações de privacidade, enquanto mecanismos de preservação da privacidade podem introduzir novos compromissos de segurança. Essa interdependência reforça a necessidade de analisar riscos de privacidade e segurança da informação de forma conjunta, e não como preocupações isoladas.

4.2 RQ2. Percepções dos Profissionais sobre Riscos de Privacidade e Segurança da Informação

Perfil dos Respondentes. Um total de 101 profissionais que atuam com privacidade e segurança da informação participou do survey. A Tabela 3 resume o perfil dos respondentes e sua experiência com ferramentas de IA e GenAI (Q1–Q5).

A maioria dos respondentes atua (Q1) na Administração Pública Federal brasileira (70,3%), seguida por empresas privadas (22,8%) e pela Administração Pública Estadual (6,9%). Essa distribuição sugere que a amostra está fortemente ancorada no contexto do setor público federal, embora ainda capture perspectivas do setor privado. Os profissionais reportam um nível relativamente elevado de formação acadêmica (Q2). A categoria mais frequente é Pós-graduação lato sensu / Especialização (34,7%), seguida por Mestrado completo (20,8%). Além disso, uma parcela substancial está atualmente cursando pós-graduação (Mestrando: 11,9%; Doutorando: 9,9%), e 9,9% possuem Doutorado completo. Um grupo menor reportou apenas graduação (12,9%). A amostra, portanto, reflete um perfil compatível com funções técnicas e de governança que frequentemente demandam formação avançada.

A área de atuação profissional predominante (Q3) é Segurança da Informação/Cibersegurança (31,3%), seguida por Desenvolvimento de Software, Sistemas e Engenharia de Requisitos (21,1%), indicando forte participação de profissionais diretamente envolvidos no desenho, implementação e proteção de sistemas de IA e GenAI. Funções orientadas à governança também estão bem representadas, com Governança de TI/Gestão de TI e Gestão/Alta Gestão somando conjuntamente 12,8% da amostra. Funções técnicas relacionadas a dados e desenvolvimento de IA (Ciência de Dados, Engenharia de Dados e IA/ML) representam 10,5%, enquanto profissionais focados em Privacidade e Proteção de Dados correspondem a 9,5%. Papéis ligados à infraestrutura totalizam 8,5% dos respondentes, e cargos de Gestão de Projetos ou Produtos representam 6,3%. Assim, a amostra incorpora perspectivas tanto operacionais quanto orientadas à governança, fortalecendo a análise ao incluir profissionais envolvidos com implementação técnica, gestão de riscos e supervisão organizacional de sistemas de IA e GenAI.

A maioria dos respondentes reporta experiência com IA ou sistemas orientados a dados (Q4) entre 1 e 3 anos (53,5%), enquanto 17,8% possuem menos de 1 ano de experiência. Entre os grupos mais experientes, 8,9% têm entre 4 e 6 anos, 4,9% entre 7 e 9 anos, e 9,9% mais de 15 anos. Proporções menores reportaram entre 10 e 12 anos (2,0%) e entre 13 e 15 anos (3,0%). De forma geral, a amostra é predominantemente composta por profissionais com experiência inicial a intermediária em atividades relacionadas à IA, com um segmento menor de profissionais altamente experientes. A maioria dos profissionais também reporta uso ou gestão ativa de ferramentas de GenAI (Q5): 50,5% indicaram uso frequente e 28,7% uso ocasional. Em contraste, 20,8% reportaram não utilizar ferramentas GenAI, trabalhando apenas com IA tradicional. Esse resultado sugere que a GenAI já constitui um elemento comum nas atividades profissionais dos respondentes, o que reforça a relevância de suas percepções sobre riscos de privacidade e segurança da informação.

Tipo de organização (Q1)	**% de respondentes**
Administração Pública Federal	70.3
Administração Pública Estadual	6.9
Empresa privada	22.8
Nível de escolaridade (Q2)	**% de respondentes**
Graduação	12.9
Pós-graduação (Especialização)	34.7
Mestrando (11.9) e Mestre (20.8)	32.7
Doutorando (9.9) e Doutor (9.9)	19.8
Área de atuação profissional (Q3)	**% de respondentes**
Segurança da Informação/Cibersegurança	31.3
Software/Desenvolvimento de Sistemas/Requisitos	21.1
Governança de TI/Gestão de TI (6.4) e Gestão/Alta Gestão (6.4)	12.8
Ciência de Dados/Engenharia de Dados/IA/ML	10.5
Privacidade e Proteção de Dados (LGPD, DPO, apoio ao DPO)	9.5
Infraestrutura de TI/Redes/Nuvem/Suporte Técnico/Operações de TI	8.5
Gestão de Projetos ou Produtos (TI)	6.3
Experiência com IA ou sistemas orientados a dados (Q4)	**% de respondentes**
Menos de 1 ano (17.8); Entre 1 e 3 anos (53.5)	71.3
Entre 4 e 6 anos	8.9
Entre 7 e 9 anos (4.9); Entre 10 e 12 anos (2.0)	6.9
Entre 13 e 15 anos (3.0); Mais de 15 anos (9.9)	12.9
Uso de ferramentas GenAI (Q5)	**% de respondentes**
Sim, frequentemente	50.5
Sim, ocasionalmente	28.7
Não (apenas IA tradicional)	20.8

Tabela 3: Perfil dos respondentes e experiência com IA e GenAI.

As Figuras 2 e 3 apresentam as percepções dos profissionais sobre a importância dos riscos de privacidade e segurança da informação associados ao uso de sistemas de Inteligência Artificial (IA) e GenAI, com base nas respostas de 101 profissionais. Em relação aos riscos de privacidade (Q6), os resultados indicam um nível muito elevado de preocupação, com a maioria dos itens sendo classificada como importante ou muito importante por uma parcela substancial dos respondentes. O vazamento de dados pessoais ou sensíveis destaca-se como o risco mais crítico: 80,0% dos respondentes o classificaram como muito importante, e outros 12,6% como importante, totalizando mais de 92% atribuindo alta relevância a esse risco. Esse achado evidencia forte consciência, por parte dos profissionais, sobre as consequências potenciais de divulgações não autorizadas no uso de IA e GenAI.

A não conformidade legal com regulações de proteção de dados (por exemplo, LGPD) também recebeu avaliações muito altas de importância, com 76,8% considerando esse risco muito importante e 13,7% importante. De modo semelhante, a falta de transparência sobre como os dados são utilizados pelos sistemas de IA foi percebida como uma preocupação crítica, com 64,2% classificando-a como muito importante e 26,3% como importante. Esses resultados enfatizam que os riscos de privacidade não são vistos apenas sob uma perspectiva técnica, mas também sob as perspectivas regulatória e de governança.

As preocupações relacionadas à inferência de atributos a partir de dados não sensíveis apresentaram um padrão um pouco mais distribuído. Embora 36,8% tenham classificado esse risco como muito importante e 45,3% como importante, uma parcela não desprezível dos respondentes o classificou como moderadamente importante (13,7%). Isso sugere que, embora a inferência de atributos seja amplamente reconhecida como uma ameaça à privacidade, sua severidade percebida varia de acordo com o contexto e o caso de uso. Por fim, a memorização de dados de treinamento por modelos de IA ou GenAI apresentou a distribuição mais equilibrada entre os riscos de privacidade. Enquanto um terço dos respondentes o classificou como muito importante (33,6%) e outro terço como importante (33,6%), uma proporção relativamente maior o considerou moderadamente importante (25,4%). Esse padrão indica consciência sobre os riscos de memorização, mas também reflete incerteza quanto à sua manifestação prática e impacto em implantações do mundo real.

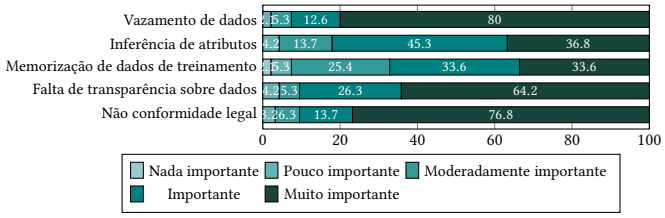

Figura 2: Importância percebida pelos profissionais dos riscos de privacidade no uso de IA e GenAI (Q6), em porcentagens (n=101).

Como mostrado na Figura 3, os riscos de segurança da informação (Q7) relacionados a sistemas de IA e GenAI também são percebidos como altamente importantes, com forte concentração de respostas nas categorias importante e muito importante em todos os itens. O uso malicioso de conteúdo gerado por IA emerge como a preocupação de segurança mais crítica, com 72,6% dos respondentes classificando-o como muito importante e 20,0% como importante. Isso reflete crescente apreensão em relação aos impactos a jusante das saídas geradas por IA, incluindo desinformação, fraude e ações automatizadas danosas.

Prompt injection e ataques indiretos por *prompt* também são percebidos como riscos de alta severidade: 69,5% dos respondentes os classificaram como muito importantes, e 22,1% como importantes. Esse resultado reforça a consciência dos profissionais sobre a interação baseada em *prompts* como uma superfície nova e significativa de ataque introduzida por sistemas GenAI. Elevada importância também foi atribuída a ataques de envenenamento de dados, com 66,4% classificando-os como muito importantes e 22,1% como importantes, bem como ao acesso não autorizado a saídas geradas por IA, classificado como muito importante por 63,2% dos respondentes. Esses achados evidenciam preocupações relacionadas tanto a ameaças em tempo de treinamento quanto em tempo de inferência. Ataques de extração ou inversão de modelos receberam avaliações ligeiramente menores, embora ainda substanciais. Enquanto 60,0% consideraram esse risco muito importante e 29,5% importante, uma pequena parcela dos respondentes o classificou como moderadamente importante (7,4%). Isso sugere que, embora a extração de modelos seja reconhecida como uma ameaça relevante, sua probabilidade ou impacto percebidos podem ser menos tangíveis para alguns profissionais quando comparados a riscos mais visíveis, como abuso de *prompts* ou uso malicioso de conteúdo.

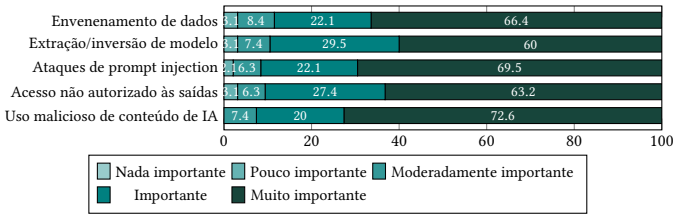

Figura 3: Importância percebida pelos profissionais dos riscos de segurança no uso de IA e GenAI (Q7), em porcentagens (n=101).

Desafios, lacunas e limitações percebidos. A Figura 4 resume as percepções dos profissionais sobre a importância dos principais desafios e lacunas na gestão de riscos de privacidade e segurança da informação no uso de sistemas de IA e GenAI (Q10). Os resultados indicam que os desafios identificados são percebidos como altamente críticos, com a maioria dos respondentes classificando-os como importantes ou muito importantes. Os desafios mais proeminentes estão relacionados a aspectos estruturais e organizacionais, e não a questões técnicas isoladas.

A falta de conhecimento técnico ou treinamento emerge como uma preocupação central, com 94,7% dos respondentes classificando-a como importante ou muito importante, e 66,3% atribuindo-lhe o nível máximo de importância. Esse achado evidencia lacunas persistentes de capacitação na compreensão e gestão dos riscos de privacidade e segurança da informação associados à IA e à GenAI. De modo semelhante, a ausência de regulações ou orientações organizacionais claras e específicas para IA e GenAI é percebida como um desafio crítico. Um total de 90,5% dos profissionais classifica essa questão como importante ou muito importante, e dois terços dos respondentes (66,3%) a consideram muito importante. Esse resultado reflete incertezas na tradução de princípios legais e éticos de alto nível em práticas operacionais concretas.

A dificuldade de equilibrar requisitos de privacidade e segurança da informação com desempenho, inovação e produtividade apresenta uma percepção mais distribuída. Embora 78,9% dos respondentes classifiquem esse desafio como importante ou muito importante, uma parcela relevante (15,8%) o considera moderadamente importante, sugerindo que os *trade-offs* entre mitigação de riscos e efetividade dos sistemas dependem do contexto e variam

entre ambientes organizacionais. A baixa maturidade organizacional ou a ausência de estruturas de governança para o uso de IA também foi identificada como uma barreira significativa. Um total de 91,6% dos respondentes percebe esse desafio como importante ou muito importante, com 61,1% atribuindo-lhe o maior nível de importância. Isso reforça o papel da governança, da responsabilização e das estruturas de tomada de decisão na gestão efetiva de riscos. A falta de transparência por parte dos fornecedores de soluções de IA e GenAI é percebida como um desafio externo crítico. Aproximadamente 89,5% dos respondentes classificam essa questão como importante ou muito importante, enfatizando a dependência das organizações em relação às informações fornecidas pelos vendedores sobre uso de dados, comportamento dos modelos e controles de risco. Esses achados indicam que os profissionais percebem as principais limitações na gestão de riscos de privacidade e segurança da informação em sistemas de IA e GenAI como predominantemente organizacionais, regulatórias e relacionadas ao conhecimento. Esses desafios ajudam a explicar a lacuna observada entre a ampla adoção de estratégias de mitigação e sua efetividade percebida limitada, conforme reportado nos resultados das Q8 e Q9.

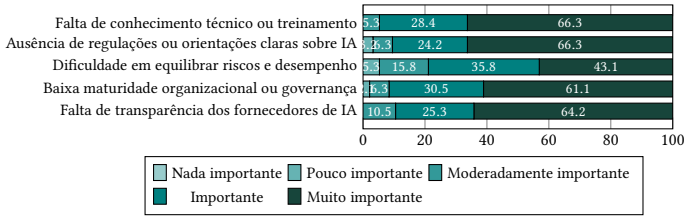

Figura 4: Importância percebida pelos profissionais dos desafios e lacunas na gestão de riscos de privacidade e segurança em IA e GenAI (Q10), em porcentagens (n=101).

A Tabela 4 resume a codificação qualitativa das respostas abertas relacionadas aos desafios de privacidade no uso de sistemas de IA e GenAI (Q11). A análise seguiu um processo de codificação indutiva, permitindo que múltiplas categorias fossem atribuídas a uma mesma resposta quando aplicável. Os achados mostram que as preocupações dos profissionais não se limitam a questões técnicas isoladas, mas são moldadas por uma combinação de comportamento humano, práticas organizacionais, arranjos de governança e fatores relacionados a fornecedores, os quais, em conjunto, influenciam os riscos de privacidade na prática.

Um dos temas mais recorrentes diz respeito ao tratamento inadequado de informações sensíveis, especialmente a inserção de informações pessoais, institucionais ou confidenciais em *prompts* de GenAI. Os participantes relataram com frequência situações em que funcionários copiam documentos internos, dados de clientes ou informações pessoais para ferramentas externas de IA sem uma compreensão clara das implicações de privacidade associadas. Práticas semelhantes também foram relatadas em contextos acadêmicos e administrativos, nos quais dados reais institucionais ou de cidadãos são utilizados sem anonimização adequada, levantando preocupações sobre exposição indevida, perda de confidencialidade e não conformidade regulatória.

Outra categoria proeminente refere-se à insuficiente conscientização e treinamento dos usuários. As respostas indicam que os usuários frequentemente subestimam os riscos de privacidade e segurança da informação ao interagir com ferramentas de IA, mesmo quando existem diretrizes formais, e que as iniciativas atuais de conscientização são limitadas e altamente dependentes do comportamento individual. A ausência frequente de treinamentos práticos e baseados em cenários evidencia uma lacuna entre políticas de alto nível e práticas operacionais cotidianas, sugerindo que os riscos de privacidade são geralmente impulsionados pelo trabalho rotineiro e pela baixa literacia em riscos, e não por intenção maliciosa. A análise também revela a prevalência de *shadow IT* e do uso não regulado de IA, especialmente em contextos do setor público. Os respondentes relataram a adoção de ferramentas GenAI não aprovadas e manifestaram preocupações sobre a integração de capacidades de IA em ambientes de desenvolvimento, como IDEs, que podem fornecer acesso a código-fonte, bases de dados ou outros ativos sensíveis sem supervisão organizacional adequada ou transparência quanto ao tratamento dos dados pelos fornecedores.

Vazamento de dados e violações de confidencialidade constituem outro tema central. Os participantes expressaram preocupações sobre a exposição de dados organizacionais estratégicos, confidenciais ou sigilosos, bem como sobre o vazamento de código-fonte, credenciais e propriedade intelectual. Esses riscos foram associados tanto às fases de treinamento quanto de inferência, reforçando a percepção de que os sistemas GenAI podem amplificar vulnerabilidades já existentes de confidencialidade. Intimamente relacionada a essa questão, a transparência do fornecedor e a custódia dos dados foram reiteradamente mencionadas como fatores estruturais de risco, com os respondentes reportando incerteza sobre como os provedores de IA armazenam, reutilizam ou compartilham dados submetidos, além de pouca visibilidade sobre os fluxos de dados uma vez que a informação deixa ambientes institucionais controlados.

Lacunas de governança e regulação emergiram como preocupações estruturais em múltiplas respostas, particularmente a ausência de políticas internas, diretrizes claras e mecanismos de responsabilização ou auditoria para o uso de GenAI. Os participantes também levantaram preocupações sobre riscos de inferência, perfilamento e vigilância, observando que interações com LLMs podem permitir inferências sensíveis sobre comportamento e usos preditivos eticamente problemáticos. Além disso, os respondentes enfatizaram riscos relacionados à confiabilidade e responsabilização das saídas geradas por IA, especialmente em contextos de alto impacto, nos quais a limitada auditabilidade e rastreabilidade podem comprometer a responsabilidade técnica e jurídica. Esses problemas são agravados por pressões organizacionais por rápida adoção da IA e por baixa prontidão institucional. Os achados qualitativos reforçam os resultados quantitativos ao mostrar que os riscos de privacidade em sistemas de IA e GenAI são percebidos como inerentemente sociotécnicos, emergindo da interação entre práticas cotidianas, lacunas de governança, ecossistemas de fornecedores e pressões organizacionais, e destacando a necessidade de abordagens integradas e orientadas à maturidade para gestão de riscos.

A análise qualitativa das respostas abertas sobre desafios de segurança relacionados ao uso de IA e GenAI (Q12) revela forte preocupação com a exposição de ativos técnicos sensíveis e com o tratamento inseguro de informações confidenciais. Os profissionais relataram com frequência vazamento acidental de credenciais, *tokens*, código-fonte, detalhes de infraestrutura e dados de acesso

Tabela 4: Codificação qualitativa das respostas abertas sobre desafios de privacidade no uso de IA e GenAI (Q11)

Categoria	Subcategoria	Citações representativas
Uso inadequado de dados sensíveis	Inserção de dados pessoais ou sensíveis em *prompts*	"Uma situação que acontece com frequência é quando funcionários copiam informações internas da empresa ou dados de clientes e colam em ferramentas gratuitas de IA para resumir ou escrever e-mails."
	Uso de dados reais em contextos acadêmicos ou administrativos	"Usar GenAI com dados reais de estudantes pode levar à exposição indevida de informações pessoais."
Falta de conscientização e treinamento dos usuários	Baixa consciência de risco entre os funcionários	"Mesmo com diretrizes, é difícil conscientizar; os controles dependem da boa vontade dos usuários."
	Ausência de treinamento prático	"Falta treinamento e exemplos práticos de perda de privacidade no dia a dia do desenvolvimento de software."
Shadow IT e uso não regulado de IA	Uso de ferramentas de IA não autorizadas	"Servidores públicos estão usando ferramentas GenAI no trabalho sem aprovação ou regulamentação interna."
	Integração de IA em IDEs e fluxos de trabalho	"IDEs integradas com IA podem acessar bases de código e possivelmente dados confidenciais, e não sabemos como os fornecedores usam essas informações."
Vazamento de dados e violações de confidencialidade	Vazamento de dados institucionais ou estratégicos	"A maior preocupação é a privacidade organizacional, como vazamento de dados sensíveis ou seu uso indevido para treinamento."
	Exposição de código-fonte e propriedade intelectual	"Vazamento de código-fonte quando desenvolvedores consultam soluções de IA durante o desenvolvimento de aplicações."
Transparência do fornecedor e custódia dos dados	Incerteza sobre o uso de dados pelos provedores de IA	"Usamos um provedor de LLM sem garantias sobre o destino dos dados lidos para sumarização."
	Hospedagem externa e soberania dos dados	"A maioria das soluções GenAI não permite custódia interna dos dados, o que impacta diretamente a conformidade com a LGPD."
Lacunas de governança e regulação	Ausência de políticas ou normas internas	"Não existe governança interna nem regulamentação para o uso de GenAI."
	Necessidade de diretrizes formais e responsabilização	"Criar uma regulamentação interna para GenAI com foco em privacidade é essencial."
Riscos de inferência, perfilamento e vigilância	Perfilamento de usuários e inferência comportamental	"Há preocupação com a criação de perfis de usuários a partir de interações com LLMs, permitindo inferências sensíveis."
	Uso preditivo ou semelhante à vigilância	"Identificação de padrões pessoais que poderiam gerar culpabilidade, algo semelhante a Minority Report."
Confiabilidade, integridade e responsabilização	Alucinações e saídas incorretas	"A informação fornecida deve estar correta e não induzir os usuários ao erro, especialmente em contextos de alto impacto."
	Auditabilidade e responsabilidade	"Logs de auditoria são necessários para permitir responsabilização técnica e jurídica caso ocorram danos."
Pressões culturais e organizacionais	Pressão por rápida adoção da IA	"Todas as áreas estão pressionadas a usar GenAI sem treinamento adequado sobre os riscos associados."
	Fraca cultura de segurança	"Há uma ausência total de cultura organizacional de segurança."

restrito ao interagir com ferramentas GenAI, especialmente durante atividades rotineiras como desenvolvimento de software, administração de sistemas, depuração e resolução de problemas. Esses riscos são ainda amplificados pelo uso de ferramentas de *shadow AI* e agentes integrados a IDEs, que podem acessar repositórios, arquivos de configuração ou ambientes de execução sem visibilidade ou controle adequados. Como mostrado na Tabela 5, tais práticas foram percebidas como ameaças críticas à segurança, pois podem viabilizar acesso não autorizado, comprometer a integridade dos sistemas e facilitar ataques subsequentes contra ativos organizacionais.

Outro tema proeminente diz respeito à crescente sofisticação das ameaças de segurança viabilizadas pela GenAI e à insuficiente conscientização dos usuários. Os respondentes destacaram o uso da IA para apoiar *phishing*, desenvolvimento de malware, fraudes, *deepfakes* e ataques de engenharia social, enfatizando que o conteúdo gerado por IA tornou-se cada vez mais realista e difícil de detectar. Em paralelo, vários participantes apontaram uma dependência excessiva das saídas da IA e uma falsa percepção de que sistemas de IA podem resolver autonomamente problemas técnicos complexos, levando à validação insuficiente de código gerado, recomendações de segurança e saídas analíticas. Essa questão também foi observada em contextos educacionais, nos quais estudantes podem gerar código inseguro ou conteúdo não confiável sem compreender suas implicações de segurança, reforçando o papel dos fatores humanos na amplificação dos riscos.

Os achados também evidenciam desafios estruturais e organizacionais relacionados à governança, à gestão de dados e à responsabilização. Muitos respondentes apontaram a ausência de políticas de segurança específicas para IA, diretrizes formais e mecanismos de monitoramento, combinadas com forte pressão organizacional por uma adoção rápida e, por vezes, descontrolada da IA. Também foram levantadas preocupações sobre práticas inadequadas de classificação de dados, desenvolvimento por terceiros, hospedagem externa, *guardrails* fracos, transparência limitada por parte dos fornecedores de IA e insuficiente auditabilidade dos sistemas. Em conjunto, esses aspectos indicam que os riscos de segurança da informação associados à IA e à GenAI não são puramente técnicos, mas estão profundamente entrelaçados com lacunas de governança, cultura organizacional, comportamento dos usuários e um cenário de ameaças em evolução moldado por ataques potencializados por IA.

4.3 RQ3. Estratégias de Mitigação Adotadas e sua Efetividade

As Figuras 5 e 6 apresentam as percepções dos profissionais quanto à adoção e à efetividade das estratégias de mitigação para riscos de privacidade e segurança da informação em sistemas de IA e GenAI (Q8 e Q9). Em relação à adoção das estratégias de mitigação (Q8), os resultados indicam forte ênfase em práticas preventivas e operacionais. Evitar o uso de dados pessoais ou sensíveis sempre que

Tabela 5: Codificação qualitativa das respostas abertas sobre desafios de segurança no uso de IA e GenAI (Q12)

Categoria	Subcategoria	Citações representativas
Exposição de ativos técnicos sensíveis	Vazamento de credenciais e segredos	"Pessoas vazando chaves, tokens e senhas em conversas com IA."
	Exposição de código-fonte e configuração de sistemas	"Desenvolvedores colam código do sistema e regras de negócio em ferramentas gratuitas de IA, levando a falhas de segurança."
Uso não regulado e inseguro de IA	Uso de ferramentas de IA não aprovadas	"Usuários empregam ferramentas GenAI não homologadas pela instituição."
	Shadow AI e agentes integrados a IDEs	"Agentes de IA integrados a IDEs aumentam riscos ao capturar tokens e credenciais do código-fonte."
Conscientização do usuário e dependência excessiva da IA	Falta de consciência sobre segurança	"O maior desafio é a conscientização dos usuários e o uso de ferramentas não homologadas."
	Ilusão de automação total	"Existe uma falsa ilusão de que a IA resolverá todos os problemas de codificação sem interferência humana."
Ameaças avançadas viabilizadas por IA	Ciberataques apoiados por IA	"A GenAI possibilita mensagens de phishing extremamente convincentes e difíceis de detectar."
	Deepfakes e engenharia social	"Deep fakes de reuniões gravadas são uma preocupação crescente."
Prompt Injection e exploração de modelos	Ataques de *prompt injection*	"Há preocupação com vulnerabilidade a ataques de prompt injection."
	Uso indevido e exploração de modelos	"O uso de algoritmos de IA para técnicas de invasão é uma preocupação séria."
Lacunas de governança, regulação e organização	Ausência de políticas institucionais de segurança	"Não existe instrução normativa regulando o uso de IA na instituição."
	Pressão por adoção descontrolada da IA	"Há forte pressão da alta gestão para adotar IA sem controles adequados."
Riscos de governança de dados e infraestrutura	Má classificação e tratamento de dados	"O uso de IA sem classificação adequada dos dados cria riscos de segurança."
	Hospedagem externa e desenvolvimento por terceiros	"Contratar terceiros para desenvolver IA sem considerar a segurança institucional é extremamente grave."
Confiabilidade, integridade e responsabilização	Vulnerabilidades em modelos e agentes de IA	"Os guardrails atuais da GenAI são fracos ou inexistentes."
	Falta de monitoramento e auditabilidade	"É necessário criar mecanismos de monitoramento e observabilidade para sistemas de IA."
Riscos éticos e de propriedade intelectual	Plágio e uso indevido de conteúdo gerado por IA	"Usar pesquisa gerada por IA sem verificar fontes pode levar a crimes de plágio."
	Apropriação indevida de propriedade intelectual	"Apropriação intelectual inadequada de dados de pesquisa acadêmica."

possível é a estratégia mais amplamente adotada, com 87,3% dos respondentes concordando ou concordando totalmente (74,7% concordam totalmente). De forma semelhante, a aplicação de técnicas de anonimização ou pseudonimização é reportada pela maioria dos profissionais, com 78,8% indicando concordância ou concordância total, embora uma parcela não desprezível permaneça neutra ou discorde, sugerindo adoção desigual entre organizações.

Medidas em nível organizacional apresentam adoção mais heterogênea. Enquanto 73,7% dos respondentes concordam ou concordam totalmente que suas organizações possuem políticas, normas ou diretrizes para o uso de IA e GenAI, 25,3% reportam neutralidade ou discordância, indicando lacunas nas estruturas formais de governança. Padrão semelhante é observado para iniciativas de treinamento e conscientização: embora 72,6% concordem ou concordem totalmente que há oferta de treinamento sobre privacidade, segurança e uso responsável da IA, quase um quarto dos respondentes reporta neutralidade ou discordância. O uso de controles técnicos de segurança e privacidade segue a mesma tendência, com 72,6% reportando concordância ou concordância total, mas 27,4% indicando implementação limitada ou inexistente.

Ao avaliar a efetividade percebida das estratégias atuais de mitigação (Q9), os profissionais expressam visões mais críticas. Para mitigação de riscos de privacidade, apenas 37,9% concordam ou concordam totalmente que as estratégias existentes são suficientes, enquanto 49,4% discordam ou discordam totalmente e 12,6% permanecem neutros. Um padrão comparável é observado para a mitigação de riscos de segurança, em que 39,0% concordam ou concordam totalmente com a suficiência das estratégias atuais, ao passo que 46,3% discordam ou discordam totalmente. Esses achados sugerem que, embora estratégias de mitigação para riscos de privacidade e segurança da informação em IA e GenAI sejam amplamente adotadas na prática, sua efetividade percebida permanece limitada. Essa lacuna evidencia um descompasso entre a presença de controles e a confiança dos profissionais em sua capacidade de mitigar adequadamente os riscos, reforçando a necessidade de abordagens mais sistemáticas, integradas e orientadas à maturidade para a gestão de riscos.

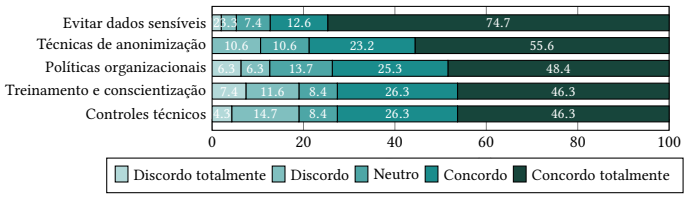

Figura 5: Concordância dos profissionais com as estratégias de mitigação adotadas para riscos de privacidade e segurança em IA e GenAI (Q8), em porcentagens (n=101).

Estratégias de mitigação reportadas na literatura. Estudos anteriores reportam um amplo conjunto de estratégias de mitigação voltadas à redução de riscos de privacidade e segurança da informação em sistemas de IA. Em nível técnico, abordagens frequentemente citadas incluem minimização de dados, técnicas de

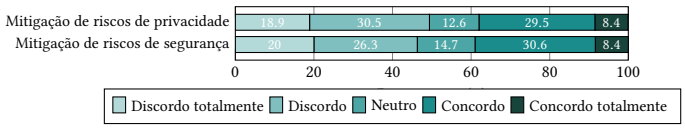

Figura 6: Efetividade percebida pelos profissionais das estratégias atuais de mitigação para riscos de privacidade e segurança em IA e GenAI (Q9), em porcentagens (n=101).

anonimização e pseudonimização, privacidade diferencial, treinamento seguro de modelos, mecanismos de controle de acesso e monitoramento de entradas e saídas dos modelos [20, 29]. No entanto, a literatura também enfatiza que essas técnicas frequentemente envolvem *trade-offs* entre privacidade, segurança da informação, utilidade do modelo e desempenho.

Para sistemas GenAI, trabalhos recentes propõem defesas em tempo de inferência, como sanitização de *prompts*, filtragem de saídas e detecção de padrões maliciosos em *prompts*, com o objetivo de mitigar ataques baseados em *prompt* e vazamentos de informação [7]. Embora essas abordagens mostrem potencial, a maioria dos estudos reconhece que sua efetividade em implantações reais de larga escala ainda permanece incerta.

Além dos controles técnicos, diversos estudos ressaltam a importância de estratégias organizacionais e orientadas à governança. Entre elas, destacam-se a adoção de princípios de privacidade desde a concepção (*privacy by design*), o estabelecimento de políticas e diretrizes claras para o uso de IA, processos de avaliação de riscos ao longo do ciclo de vida da IA e o alinhamento com *frameworks* de gestão de riscos em IA [26]. Ainda assim, a literatura aponta uma lacuna persistente entre a disponibilidade desses *frameworks* e sua implementação prática, especialmente em contextos que envolvem GenAI. Os resultados da literatura revelam que, embora um conjunto diversificado de riscos de privacidade e segurança da informação tenha sido identificado e múltiplas estratégias de mitigação tenham sido propostas, as abordagens existentes permanecem fragmentadas. Medidas técnicas, organizacionais e regulatórias frequentemente são tratadas de forma isolada, o que limita sua efetividade na gestão combinada dos riscos associados a sistemas de IA e GenAI.

5 DISCUSSÃO

Embora diversos estudos anteriores identifiquem categorias semelhantes de riscos de privacidade e segurança da informação, este estudo contribui empiricamente ao revelar como esses riscos se manifestam nas práticas organizacionais cotidianas, além de evidenciar uma lacuna sistemática entre a ampla adoção de estratégias de mitigação e a confiança dos profissionais em sua efetividade no uso real de sistemas GenAI. Ao integrar resultados quantitativos e qualitativos com a literatura existente, nossos achados tanto confirmam quanto ampliam pesquisas anteriores, oferecendo uma compreensão socio-técnica mais abrangente desses riscos na prática.

Primeiramente, nossos resultados corroboram fortemente estudos prévios que identificam vazamento de dados, falta de transparência e não conformidade regulatória como preocupações centrais de privacidade e segurança da informação em sistemas de IA. De forma semelhante às taxonomias de risco propostas por [29] e [20], os profissionais participantes do survey classificaram consistentemente o vazamento de dados pessoais ou informações sensíveis e a falta de transparência no uso de dados como riscos muito importantes. No entanto, nossos resultados ampliam essas contribuições ao mostrar que tais riscos não são percebidos apenas como problemas abstratos ou puramente técnicos, mas como situações concretas e recorrentes inseridas em práticas rotineiras de trabalho, como a cópia de informações sensíveis em prompts de GenAI. Essa observação reforça evidências empíricas provenientes de estudos centrados em desenvolvedores, que indicam que a conformidade com requisitos de privacidade frequentemente ocorre de forma reativa e ad hoc, em vez de ser incorporada sistematicamente aos processos de engenharia de software [16].

Em segundo lugar, o estudo confirma e aprofunda a literatura sobre riscos de segurança em sistemas GenAI, especialmente aqueles decorrentes de interações baseadas em prompts e do uso indevido de conteúdos gerados por IA. Em consonância com os modelos de ameaça descritos por [7], nossos respondentes classificaram ataques de *prompt injection* e ataques indiretos por prompt entre os riscos de segurança mais severos. As respostas qualitativas ilustram ainda como essas vulnerabilidades se manifestam na prática, incluindo a exposição de credenciais, código-fonte e detalhes de infraestrutura durante atividades rotineiras de desenvolvimento e resolução de problemas. Além disso, os profissionais destacaram o aumento da sofisticação de ataques cibernéticos habilitados por IA, como phishing automatizado, geração de malware e engenharia social, preocupação também evidenciada em estudos organizacionais focados em segurança [8]. Esses resultados reforçam que as ameaças em tempo de inferência introduzidas por sistemas GenAI representam uma mudança qualitativa no cenário de segurança da IA.

Em terceiro lugar, nossos resultados reforçam pesquisas anteriores que indicam que fatores organizacionais e humanos desempenham papel decisivo na exposição a riscos relacionados à IA. De maneira semelhante aos achados de [9] e [14], os profissionais relatam ampla adoção de estratégias básicas de mitigação, como evitar o uso de informações sensíveis e aplicar técnicas de anonimização. No entanto, nossa análise revela uma lacuna persistente entre a adoção dessas medidas e sua efetividade percebida. Essa lacuna reflete preocupações levantadas em estudos orientados à governança, os quais argumentam que estruturas e políticas de alto nível frequentemente falham em se traduzir em práticas operacionais concretas [26]. Nossos resultados demonstram empiricamente que essa desconexão não é apenas conceitual, mas vivenciada diretamente pelos profissionais, que expressam baixa confiança na suficiência dos controles atualmente adotados.

Além disso, o estudo amplia pesquisas sobre adoção organizacional de IA ao destacar o papel das pressões institucionais e estratégicas. Estudos sobre tomada de decisão orientada por IA sugerem que organizações frequentemente priorizam inovação, eficiência e ganhos de desempenho, muitas vezes avançando mais rapidamente do que a criação de estruturas adequadas de governança e gestão de riscos [10]. Nossos resultados qualitativos estão fortemente alinhados a essa perspectiva, pois diversos respondentes relataram pressão da alta gestão para adotar GenAI rapidamente, frequentemente sem políticas claras, treinamento adequado ou mecanismos de monitoramento. Essa dinâmica ajuda a explicar por que estratégias

de mitigação são amplamente reportadas, mas simultaneamente percebidas como insuficientes.

Por fim, nossos resultados contribuem para o crescente corpo de pesquisas que compreendem os riscos da IA como fenômenos inerentemente socio-técnicos. Em consonância com estudos voltados à ética e governança [24], os participantes destacaram questões como responsabilização, auditabilidade, transparência de fornecedores e custódia de dados como desafios fundamentais. Essas preocupações extrapolam o escopo tradicional da segurança e da engenharia de privacidade em software, indicando que a gestão eficaz de riscos em sistemas de IA e GenAI requer respostas coordenadas em níveis técnico, organizacional e regulatório. Ao relacionar empiricamente percepções de profissionais, incidentes do mundo real e lacunas de governança, este estudo avança a literatura existente, que frequentemente tem tratado essas dimensões de forma isolada.

Em síntese, esta pesquisa confirma evidências já existentes de que os riscos de privacidade e segurança da informação em sistemas de IA são significativos e multifacetados, ao mesmo tempo em que amplia estudos anteriores ao oferecer uma perspectiva empírica integrada baseada nas experiências de profissionais. A convergência entre nossos resultados quantitativos e qualitativos e a literatura existente reforça que a mitigação de riscos relacionados à IA exige mais do que salvaguardas técnicas: requer estruturas de governança orientadas à maturidade organizacional, capacitação contínua, maior clareza regulatória e maior transparência por parte dos fornecedores de soluções de IA.

6 AMEAÇAS À VALIDADE

Este estudo está sujeito a diversas ameaças à validade, discutidas de acordo com categorias amplamente adotadas em pesquisas empíricas em engenharia de software [33].

Validade de construto refere-se à capacidade do instrumento de survey de capturar adequadamente os conceitos de riscos de privacidade, riscos de segurança da informação, estratégias de mitigação e desafios organizacionais relacionados ao uso de IA e GenAI. Para mitigar essa ameaça, os itens do survey foram sistematicamente derivados de uma revisão abrangente da literatura sobre privacidade, segurança da informação e governança em IA. Além disso, foi conduzido um estudo piloto com profissionais que atuam com privacidade e segurança da informação para avaliar a clareza e a interpretabilidade das perguntas, resultando em pequenos ajustes de redação. Ainda assim, alguns construtos — como a efetividade percebida das estratégias de mitigação — permanecem inerentemente subjetivos e dependentes das experiências individuais e dos contextos organizacionais dos respondentes.

Validade interna apresenta ameaças limitadas, uma vez que este estudo não busca estabelecer relações causais, mas sim explorar percepções e práticas reportadas. Entretanto, dados auto-relatados podem ser influenciados por vieses individuais, como desejabilidade social ou exposição prévia a incidentes de segurança. Para reduzir esse risco, o survey foi conduzido de forma anônima e explicitamente apresentado como um instrumento para coleta de percepções, e não como avaliação do desempenho organizacional.

Validade externa refere-se à generalização dos resultados. Os participantes do survey foram recrutados por amostragem por conveniência, por meio de redes profissionais e mídias sociais, o que pode limitar a representatividade da amostra. Além disso, a amostra está fortemente concentrada no contexto do setor público brasileiro. Embora esse foco aumente a relevância dos resultados para ambientes organizacionais e regulatórios semelhantes, é necessário cautela ao generalizar os achados para outros países ou para contextos do setor privado com estruturas de governança e níveis de maturidade distintos.

Confiabilidade diz respeito à consistência e reprodutibilidade da análise qualitativa. Para mitigar essa ameaça, foi adotada uma abordagem sistemática de codificação indutiva das respostas abertas, com foco na identificação de temas recorrentes entre múltiplos participantes, em vez de declarações isoladas. Ainda assim, a interpretação qualitativa envolve necessariamente julgamento dos pesquisadores, e esquemas alternativos de codificação poderiam enfatizar diferentes aspectos dos dados. Apesar dessas limitações, a triangulação entre evidências da literatura e dados quantitativos e qualitativos do survey aumenta a robustez dos resultados e fortalece a validade das conclusões apresentadas.

7 CONCLUSÃO E TRABALHOS FUTUROS

Este artigo investigou os riscos de privacidade e segurança da informação associados ao uso de sistemas de Inteligência Artificial e GenAI por meio de uma análise integrada da literatura existente e de um survey empírico com profissionais que atuam com privacidade e segurança da informação no Brasil. Os resultados indicam que os profissionais percebem os riscos de privacidade e segurança da informação como altamente críticos e predominantemente de natureza socio-técnica. Riscos como vazamento de dados, não conformidade regulatória, falta de transparência, ataques baseados em prompts e uso malicioso de conteúdos gerados por IA foram consistentemente classificados como importantes ou muito importantes. Esses achados reforçam a visão de que os desafios de privacidade e segurança da informação em sistemas de IA vão além de vulnerabilidades técnicas isoladas, sendo profundamente influenciados por práticas organizacionais, estruturas de governança, comportamento dos usuários e ecossistemas de fornecedores.

O estudo também revela uma lacuna significativa entre a adoção de estratégias de mitigação e sua efetividade percebida. Embora práticas preventivas — como evitar o uso de informações sensíveis, aplicar técnicas de anonimização e estabelecer políticas organizacionais — sejam amplamente reportadas, os profissionais expressam confiança limitada na capacidade dessas medidas de mitigar adequadamente riscos de privacidade e segurança da informação em sistemas de IA e GenAI. Desafios relacionados à falta de treinamento, baixa maturidade organizacional, ausência de orientações claras e transparência limitada por parte dos fornecedores de IA ajudam a explicar essa discrepância e evidenciam barreiras estruturais à gestão eficaz de riscos.

Ao integrar perspectivas técnicas, organizacionais e humanas, este estudo contribui para uma compreensão mais holística dos riscos de privacidade e segurança da informação no uso de IA e GenAI. Os resultados corroboram e ampliam pesquisas anteriores ao demonstrar empiricamente que lacunas de governança, limitações de conhecimento e interações socio-técnicas desempenham papel central tanto na percepção dos riscos quanto nas práticas de mitigação adotadas.

Como direções futuras de pesquisa, diversos caminhos podem ser explorados. Primeiramente, estudos longitudinais podem investigar como as percepções de risco e as práticas de mitigação evoluem à medida que as organizações adquirem maior experiência com GenAI e à medida que os marcos regulatórios amadurecem. Em segundo lugar, estudos comparativos entre países ou setores podem fornecer insights mais aprofundados sobre a influência de ambientes regulatórios e contextos organizacionais distintos. Em terceiro lugar, pesquisas futuras podem focar no desenvolvimento e na avaliação de modelos de maturidade, diretrizes ou ferramentas de apoio à decisão que traduzam princípios de privacidade e segurança da informação em práticas operacionais aplicáveis a sistemas GenAI. Por fim, estudos experimentais e baseados em estudos de caso podem complementar os resultados de surveys ao avaliar empiricamente a efetividade real de estratégias técnicas e organizacionais de mitigação em implantações operacionais de sistemas de IA.

REFERÊNCIAS

[1] Sara Abdali, Richard Anarfi, C. J. Barberan, and Jia He. 2024. Securing Large Language Models: Threats, Vulnerabilities and Responsible Practices. *CoRR* abs/2403.12503 (2024), 1–48. arXiv:2403.12503 doi:10.48550/ARXIV.2403.12503

[2] Lakshay Batra, Shruti Batra, and Vinish Kathuria. 2025. Leveraging Human-Centered AI Framework to Mitigate Security, Privacy, and Ethical Risks in AI Agents. In *HCI International 2025 Posters - 27th International Conference on Human-Computer Interaction, HCII 2025, Gothenburg, Sweden, June 22-27, 2025, Proceedings, Part IV (Communications in Computer and Information Science, Vol. 2525)*, Constantine Stephanidis, Margherita Antona, Stavroula Ntoa, and Gavriel Salvendy (Eds.). Springer, https://doi.org/10.1007/978-3-031-94159-7_2, 10–17. doi:10.1007/978-3-031-94159-7_2

[3] Grace Billiris, Asif Gill, and Madhushi Bandara. 2025. Privacy in the Age of AI: A Taxonomy of Data Risks. *CoRR* abs/2510.02357 (2025), 1–12. arXiv:2510.02357 doi:10.48550/ARXIV.2510.02357

[4] Antony Bryant and Kathy Charmaz. 2007. *The Sage handbook of grounded theory*. Sage, https://uk.sagepub.com/en-gb/eur/the-sage-handbook-of-grounded-theory/book234413.

[5] Nartional Congress da República, Presidência. 2018. Brazilian General Data Protection Law (LGPD). *Nartional Congress, accessed in April 10, 2022* 1, 1 (2018), 1–31. https://www.pnm.adv.br/wp-content/uploads/2018/08/Brazilian-General-Data-Protection-Law.pdf

[6] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2025. Security and Privacy Challenges of Large Language Models: A Survey. *ACM Comput. Surv.* 57, 6 (2025), 152:1–152:39. doi:10.1145/3712001

[7] Erik Derner, Kristina Batistic, Jan Zahálka, and Robert Babuska. 2024. A Security Risk Taxonomy for Prompt-Based Interaction With Large Language Models. *IEEE Access* 12 (2024), 126176–126187. doi:10.1109/ACCESS.2024.3450388

[8] Fatema EL-Husseini, Hassan N Noura, and Flavien Vernier. 2025. Security and privacy-preserving for machine learning models: attacks, countermeasures, and future directions. *Annals of Telecommunications* 80 (2025), 1–22.

[9] Fabiano Damasceno Sousa Falcão and Edna Dias Canedo. 2024. Investigating Software Development Teams Members' Perceptions of Data Privacy in the Use of Large Language Models (LLMs). In *Proceedings of the XXIII Brazilian Symposium on Software Quality, SBQS 2024, Salvador, Bahia, Brazil, November 5-8, 2024*, Ivan Machado, José Carlos Maldonado, Tayana Conte, Edna Dias Canedo, Johnny Marques, Breno Bernard Nicolau de França, Patrícia Matsubara, Davi Viana, Sérgio Soares, Gleison Santos, Larissa Rocha, Bruno Gadelha, Rodrigo Pereira dos Santos, Ana Carolina Oran, and Adolfo Gustavo Serra Seca Neto (Eds.). ACM, https://doi.org/10.1145/3701625.3701675, 373–382. doi:10.1145/3701625.3701675

[10] Massimo Fascinari and Vincent English. 2025. The Impact of Artificial Intelligence on Strategic Technology Management: A Mixed-Methods Analysis of Resources, Capabilities, and Human-AI Collaboration. *arXiv preprint arXiv:2512.08938* 1 (2025), 1–32.

[11] Georgios Feretzakis, Konstantinos Papaspyridis, Aris Gkoulalas-Divanis, and Vassilios S. Verykios. 2024. Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review. *Inf.* 15, 11 (2024), 697. doi:10.3390/INFO15110697

[12] Furizal Furizal, Agus Ramelan, Feri Adriyanto, Hari Maghfiroh, Alfian Ma'arif, Kariyamin Kariyamin, Alya Masitha, and Aldi Bastiatul Fawait. 2024. Concerns of Ethical and Privacy in the Rapid Advancement of Artificial Intelligence: Directions, Challenges, and Solutions. *Journal of Robotics and Control (JRC)* 5, 6 (2024), 2015–2026.

[13] Abenezer Golda, Kidus Mekonen, Amit Pandey, Anushka Singh, Vikas Hassija, Vinay Chamola, and Biplab Sikdar. 2024. Privacy and Security Concerns in Generative AI: A Comprehensive Survey. *IEEE Access* 12 (2024), 48126–48144. doi:10.1109/ACCESS.2024.3381611

[14] Clendson Domingos Gonçalves, Eduardo de Paoli Menescal, Fábio Lúcio Lopes de Mendonça, and Edna Dias Canedo. 2024. Trust in AI: Perspectives of C-Level Executives in Brazilian Organizations. In *Proceedings of the XXIII Brazilian Symposium on Software Quality, SBQS 2024, Salvador, Bahia, Brazil, November 5-8, 2024*, Ivan Machado, José Carlos Maldonado, Tayana Conte, Edna Dias Canedo, Johnny Marques, Breno Bernard Nicolau de França, Patrícia Matsubara, Davi Viana, Sérgio Soares, Gleison Santos, Larissa Rocha, Bruno Gadelha, Rodrigo Pereira dos Santos, Ana Carolina Oran, and Adolfo Gustavo Serra Seca Neto (Eds.). ACM, https://doi.org/10.1145/3701625.3701654, 147–157. doi:10.1145/3701625.3701654

[15] Mohd Ariful Haque, Sunzida Siddique, Md. Mahfuzur Rahman, Ahmed Rafi Hasan, Laxmi Rani Das, Marufa Kamal, Tasnim Masura, and Kishor Datta Gupta. 2025. SOK: Exploring Hallucinations and Security Risks in AI-Assisted Software Development with Insights for LLM Deployment. *CoRR* abs/2502.18468 (2025), 1–27. arXiv:2502.18468 doi:10.48550/ARXIV.2502.18468

[16] Georgia M Kapitsaki, Maria Papoutsoglou, Christoph Treude, and Ioanna Theophilou. 2025. Analyzing developer discussions on EU and US privacy legislation compliance in GitHub repositories. *arXiv preprint arXiv:2512.10618* 1 (2025), 1–40.

[17] Barbara A. Kitchenham and Shari Lawrence Pfleeger. 2008. Personal Opinion Surveys. In *Guide to Advanced Empirical Software Engineering*, Forrest Shull, Janice Singer, and Dag I. K. Sjøberg (Eds.). Springer, https://doi.org/10.1007/978-1-84800-044-5_3, 63–92. doi:10.1007/978-1-84800-044-5_3

[18] Alexandra Klymenko, Stephen Meisenbacher, Patrick Gage Kelley, Sai Teja Peddinti, Kurt Thomas, and Florian Matthes. 2025. "We are not Future-ready": Understanding AI Privacy Risks and Existing Mitigation Strategies from the Perspective of AI Developers in Europe. In *Twenty-First Symposium on Usable Privacy and Security, SOUPS 2025, Seattle, WA, USA, August 10-12, 2025*, Patrick Gage Kelley, Mainack Mondal, and Kami Vaniea (Eds.). USENIX Association, https://www.usenix.org/conference/soups2025/presentation/klymenko, 113–132.

[19] Rrezarta Krasniqi, Depeng Xu, and Marco Vieira. 2026. SE Perspective on LLMs: Biases in Code Generation, Code Interpretability, and Code Security Risks. *ACM Comput. Surv.* 58, 5 (2026), 137:1–137:16. doi:10.1145/3774324

[20] Maria Lachova. 2024. Data, Privacy and Human-Centered AI in Defense and Security Systems: Legal and Ethical Considerations. *Information & Security* 55, 2 (2024), 213–221. doi:10.11610/isij.5551

[21] Donghyeok Lee, Christina Todorova, and Alireza Dehghani. 2024. Ethical Risks and Future Direction in Building Trust for Large Language Models Application under the EU AI Act. In *Proceedings of the 2024 Conference on Human Centred Artificial Intelligence - Education and Practice, HCAIep 2024, Naples, Italy, December 2-3, 2024*, Keith Nolan, Rosemary Monahan, Stefano Marrone, and Huib Aldewereld (Eds.). ACM, https://doi.org/10.1145/3701268.3701272, 41–46. doi:10.1145/3701268.3701272

[22] Yihao Liu, Jinhe Huang, Yanjie Li, Dong Wang, and Bin Xiao. 2025. Generative AI model privacy: a survey. *Artif. Intell. Rev.* 58, 1 (2025), 33. doi:10.1007/S10462-024-11024-6

[23] Zhenhua Liu, Tong Zhu, Chuanyuan Tan, and Wenliang Chen. 2025. Learning to Refuse: Towards Mitigating Privacy Risks in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, https://aclanthology.org/2025.coling-main.114/, 1683–1698.

[24] Rhodes Massenon, Ishaya Gambo, and Javed Ali Khan. 2026. Unveiling hidden permissions: an LLM framework for detecting privacy and security concerns in AI mobile apps reviews. *Automated Software Engineering* 33, 2 (2026), 42.

[25] Evgenia Novikova, Elena Doynikova, and Igor V. Kotenko. 2025. What are your privacy risks? Privacy risk assessment based on privacy policies analysis. *Expert Syst. Appl.* 280 (2025), 127270. doi:10.1016/J.ESWA.2025.127270

[26] National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). *National Institute of Standards and Technology* 1 (2023), 1–48. doi:10.6028/NIST.AI.100-1

[27] Md. Mostafizur Rahman, Aiasha Siddika Arshi, Md. Golam Moula Mehedi Hasan, Sumayia Farzana Mishu, Hossain Shahriar, and Fan Wu. 2023. Security Risk and Attacks in AI: A Survey of Security and Privacy. In *47th IEEE Annual Computers, Software, and Applications Conference, COMPSAC 2023, Torino, Italy, June 26-30, 2023*, Hossain Shahriar, Yuuichi Teranishi, Alfredo Cuzzocrea, Moushumi Sharmin, Dave Towey, A. K. M. Jahangir Alam Majumder, Hiroki Kashiwazaki, Ji-Jiang Yang, Michiharu Takemoto, Nazmus Sakib, Ryohei Banno, and Sheikh Iqbal Ahamed (Eds.). IEEE, https://doi.org/10.1109/COMPSAC57700.2023.00284, 1834–1839. doi:10.1109/COMPSAC57700.2023.00284

[28] Antonia Sattlegger and Nitesh Bharosa. 2024. Beyond principles: Embedding ethical AI risks in public sector risk management practice. In *Proceedings of the 25th Annual International Conference on Digital Government Research, DGO 2024, Taipei,*

Taiwan, June 11-14, 2024, Hsin-Chung Liao, David Duenas-Cid, Marie Anne Macadar, and Flavia Bernardini (Eds.). ACM, https://doi.org/10.1145/3657054.3657063, 70–80. doi:10.1145/3657054.3657063

[29] Sakib Shahriar, Sonal Allana, Seyed Mehdi Hazratifard, and Rozita Dara. 2023. A Survey of Privacy Risks and Mitigation Strategies in the Artificial Intelligence Life Cycle. *IEEE Access* 11 (2023), 61829–61854. doi:10.1109/ACCESS.2023.3287195

[30] Aviral Srivastava and Sourav Panda. 2024. A Formal Framework for Assessing and Mitigating Emergent Security Risks in Generative AI Models: Bridging Theory and Dynamic Risk Mitigation. *CoRR* abs/2410.13897 (2024), 1–17. arXiv:2410.13897 doi:10.48550/ARXIV.2410.13897

[31] Lukas Struppek. 2025. *Understanding and Mitigating Security, Privacy, and Ethical Risks in Generative Artificial Intelligence.* Ph. D. Dissertation. Technical University of Darmstadt, Germany, http://tuprints.ulb.tu-darmstadt.de/30054/.

[32] The European Parliament and The Council. 2018. General Data Protection Regulation (GDPR): EU Data Protection Rules. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

[33] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in Software Engineering.* Springer, https://doi.org/10.1007/978-3-642-29044-2. doi:10.1007/978-3-642-29044-2