Privacy Risks of Generative AI in Organizations and Governance Models for Risk Mitigation

Cristina Renata Allein Fontes and Edna Dias Canedo[1][a]
[1]*University of Brasília (UnB), Brasília–DF, Brazil*
cristina.fontes@gmail.com, ednacanedo@unb.br

Keywords: Generative AI, Large Language Models, Privacy Risks, Risk Governance, AI Governance, Risk Management, Privacy Impact Assessment, Public Sector

Abstract: **Context:** The adoption of Generative Artificial Intelligence (GenAI), especially Large Language Models (LLMs), is accelerating across organizational processes that handle personal, sensitive, and regulated data. Alongside productivity gains, recent evidence has raised concerns about new and evolving privacy risks that challenge traditional data protection and compliance practices. **Goal:** This paper aims to synthesize the main privacy risks associated with GenAI in organizational deployments and to identify governance-oriented mechanisms that can mitigate these risks beyond isolated technical controls, with particular attention to high-impact public-sector settings. **Method:** We conducted a structured synthesis of recent literature on privacy risks and attacks in LLM-based systems, privacy and risk assessment methodologies, and AI governance frameworks. The evidence was organized into recurrent privacy risk categories and corresponding governance mechanisms, resulting in (i) a summary of representative studies and risks and (ii) a mapping between privacy risk categories and governance-oriented mitigation strategies. **Results:** The synthesis indicates that privacy risks in GenAI are dynamic, context-dependent, and socio-technical, emerging from the interaction between models, data, users, and organizational practices. Key risk patterns include data memorization and inference-based leakage, prompt-driven disclosure and exfiltration, and lifecycle emergent risks amplified by shadow AI and fragmented accountability. Effective mitigation therefore requires integrated governance models combining technical safeguards with lifecycle-based risk management, continuous privacy assessment, clear accountability structures, and auditable regulatory alignment. We also derive practical recommendations for public-sector governance, emphasizing institutional oversight, continuous monitoring, and inter-organizational data governance to sustain privacy protection and public trust.

1 INTRODUCTION

The rapid diffusion of Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs), has significantly transformed organizational practices related to information processing, decision support, customer interaction, and the automation of knowledge intensive tasks (Ahi and Valizadeh, 2025). Enterprises across multiple sectors increasingly integrate GenAI systems into workflows that involve personal, sensitive, and sometimes regulated data, including domains such as healthcare, finance, and public services (Wang and Zare, 2023; Esposito and Tse, 2024). As a result, privacy has emerged as a critical organizational concern rather than a purely technical or regulatory issue, requiring coordinated decision-making, accountability, and governance mechanisms (Azevedo and Canedo, 2025; Sattlegger and Bharosa, 2024; Dotan et al., 2024; Matos et al., 2025; Spósito et al., 2025b). Despite growing awareness of privacy risks, many organizations continue to rely predominantly on technical safeguards such as encryption, anonymization, and access control assuming that these measures are sufficient to ensure adequate data protection in GenAI-enabled systems, an assumption that has been increasingly questioned in the literature (Chen et al., 2025; Rathod et al., 2025).

Recent research has demonstrated, however, that privacy risks in GenAI extend beyond traditional data protection challenges. Studies have shown that LLMs may unintentionally memorize personal data during training, enabling attacks such as membership inference and data reconstruction (Ruhländer et al., 2024; Chen et al., 2025). Moreover, the interaction paradigm introduced by GenAI particu-

[a] https://orcid.org/0000-0002-2159-339X

larly prompt-based interaction has created new vectors for privacy leakage, including direct and indirect prompt injection, contextual inference, and exfiltration of personal information through seemingly legitimate queries (Derner et al., 2024; Chernyshev et al., 2024). These risks are further amplified in organizational contexts, where users with varying levels of expertise interact with GenAI systems embedded in complex socio-technical environments.

While the literature on privacy and security risks in GenAI has expanded rapidly, it remains largely fragmented and predominantly technical in nature. Existing surveys and threat analyses provide detailed taxonomies of attacks, vulnerabilities, and privacy-enhancing technologies (Rathod et al., 2025; Barberá, 2025). However, less attention has been given to how these risks manifest through organizational practices, governance structures, and decision-making processes. In particular, there is limited discussion on how traditional privacy management instruments such as Privacy Impact Assessments (PIA), Data Protection Impact Assessments (DPIA), and organizational risk management frameworks can be effectively adapted to the dynamic, opaque, and probabilistic nature of GenAI systems (Wairimu et al., 2024).

This gap becomes especially problematic when considering that privacy risks in GenAI often emerge not only at design time but also during deployment and use. Empirical evidence indicates that privacy leakage may depend on contextual factors such as prompt composition, system configuration, integration with external tools, and user behavior (Chang et al., 2024; Schwartzman, 2024). Consequently, one-off assessments or compliance-oriented checklists are insufficient to capture evolving risks over the system lifecycle. Several studies argue that effective mitigation requires continuous monitoring, clearly defined roles and responsibilities, and integration of privacy considerations into broader AI governance and risk management practices (Dotan et al., 2024; Sattlegger and Bharosa, 2024).

In response to these challenges, emerging governance-oriented approaches propose viewing privacy risks in GenAI as socio-technical risks that must be addressed through a combination of technical, organizational, and procedural controls. Frameworks such as the NIST AI Risk Management Framework emphasize governance, accountability, and lifecycle-based risk management, yet they remain high-level and require contextualization for GenAI use in enterprise environments (of Standards and Technology, 2023). Similarly, recent work on algorithmic governance highlights the need to embed ethical, legal, and privacy considerations into organizational structures and decision-making processes, rather than treating them as external compliance requirements (Esposito and Tse, 2024). Despite these advances, there is still a lack of consolidated guidance that explicitly connects GenAI specific privacy risks to actionable governance mechanisms in organizational contexts.

This study addresses the following research question: **RQ. What are the main privacy risks associated with the use of Generative AI in organizations, and how can governance models mitigate these risks in enterprise contexts?** To answer this question, we conduct a structured synthesis of recent research on privacy risks, attacks, and mitigation strategies related to GenAI, with a particular focus on their implications for organizational governance and risk management. Rather than proposing new technical defenses, this study seeks to bridge the gap between technical risk identification and organizational decision-making.

The contributions of this paper are threefold. First, we consolidate and systematize the main privacy risks associated with GenAI from an organizational perspective, highlighting how these risks emerge across different stages of the AI lifecycle and usage contexts. Second, we critically discuss the limitations of purely technical mitigation strategies when applied in enterprise environments. Third, we analyze how governance-oriented mechanisms including risk management frameworks, privacy impact assessments, and algorithmic governance practices can support more effective, accountable, and sustainable mitigation of privacy risks in organizational deployments of Generative AI.

2 RELATED WORK

The governance of privacy in organizational contexts has traditionally been shaped by data protection regulations and international guidelines that establish principles such as lawfulness, purpose limitation, data minimization, transparency, and accountability. Regulatory frameworks including the General Data Protection Regulation (GDPR) in the European Union (Parliament and Council, 2018), the Brazilian General Data Protection Law (LGPD) (da República, 2018), the California Consumer Privacy Act (CCPA) (of California Department of Justice, 2018), the Australian Privacy Act (OAIC, 1988), and the proposed American Data Privacy and Protection Act (ADPPA) (of the Whole House on the State of the Union, 2022) provide legal obligations for organizations processing personal data. Complementary guidance from international bodies, such as the OECD Privacy Guide-

lines, further emphasizes organizational responsibility and cross-border cooperation in data governance (for Economic Co-operation and (OECD), 2024b).

With the increasing adoption of Artificial Intelligence systems, these regulatory instruments have been progressively extended to address AI specific risks. Recent initiatives, such as the European Union Artificial Intelligence Act (Sonsini and Parliament, 2024) and national AI governance strategies, highlight the need to integrate privacy protection into broader AI governance frameworks. Studies comparing data protection laws and privacy frameworks indicate that, despite differences in scope and enforcement mechanisms, these regulations converge in treating privacy as an organizational governance challenge rather than a purely technical compliance requirement (Rocha and Canedo, 2025). However, existing regulatory instruments were not originally designed to address the adaptive, probabilistic, and opaque characteristics of Generative AI systems, which complicates their direct application in practice.

Privacy and Security Risks in Generative AI and Large Language Models. A growing body of research has investigated privacy and security risks associated with Generative AI, particularly Large Language Models. Empirical studies and surveys identify a wide range of risks, including unintended memorization of training data, membership inference attacks, model inversion, and leakage of personal information through model outputs (Chen et al., 2025; Ruhländer et al., 2024; Rathod et al., 2025). These risks are exacerbated by the prompt-based interaction paradigm, which introduces new attack surfaces and enables both direct and indirect prompt injection attacks capable of exfiltrating sensitive information (Derner et al., 2024; Chernyshev et al., 2024; Schwartzman, 2024).

Recent surveys emphasize that privacy risks in LLMs are not confined to the training phase but may emerge dynamically during deployment and use, depending on contextual factors such as prompt composition, system configuration, and integration with external tools or data sources (Chang et al., 2024). Moreover, the dual-use nature of GenAI has been highlighted in cybersecurity research, showing that the same capabilities that support defensive applications can also be exploited to automate attacks and amplify privacy threats (Ahi and Valizadeh, 2025). These findings suggest that technical mitigation mechanisms alone are insufficient to address the evolving risk landscape of GenAI in organizational environments.

Privacy Risk Assessment and Impact Assessment Approaches Privacy Impact Assessment (PIA) and Privacy Risk Assessment (PRA) have long been used as instruments to support compliance and risk management in data processing activities. However, recent systematic reviews indicate that existing PIA and PRA methodologies suffer from limited empirical validation, high subjectivity, and weak integration with organizational decision-making processes (Wairimu et al., 2024). These limitations become particularly pronounced in the context of Generative AI systems, where risk factors may evolve over time and emerge through interaction rather than static system design.

The adaptive and probabilistic behavior of LLMs challenges traditional assessment approaches that assume stable system boundaries and predictable data flows. As a result, several authors argue for continuous and lifecycle-oriented risk assessment practices that extend beyond one-off compliance exercises (Dotan et al., 2024). Such approaches align with emerging views that privacy risks in GenAI should be managed as ongoing organizational risks rather than isolated technical vulnerabilities.

AI Governance and Risk Management Frameworks In response to the growing complexity of AI related risks, governance-oriented frameworks have been proposed to support organizational oversight and accountability. The NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0) provides a structured approach to identifying, assessing, and managing AI risks across the system lifecycle, emphasizing governance, accountability, and continuous monitoring (of Standards and Technology, 2023). Building on this foundation, recent work proposes maturity models to operationalize the AI RMF in organizational contexts, highlighting the need for coordinated roles, processes, and metrics to manage AI risks effectively (Dotan et al., 2024).

Complementary research on algorithmic governance further stresses the importance of embedding ethical, legal, and privacy considerations into organizational risk management practices. Rather than treating privacy as an external compliance requirement, these approaches advocate for integrating privacy risk management into strategic and operational decision-making structures (Sattlegger and Bharosa, 2024; Esposito and Tse, 2024). Despite these advances, there remains a lack of consolidated guidance that explicitly connects GenAI-specific privacy risks to actionable governance mechanisms in enterprise settings, motivating the need for further synthesis and analysis.

3 RESEARCH APPROACH

This study adopts a structured literature synthesis approach to analyze privacy risks associated with Generative Artificial Intelligence (GenAI), with particular emphasis on Large Language Models (LLMs), and to identify governance-oriented mechanisms capable of mitigating these risks in organizational contexts. Rather than conducting a full systematic literature review, the study follows established principles of rigor and transparency commonly recommended for structured and narrative literature reviews in Information Systems and Software Engineering research (Pautasso, 2019).

The overall research process is illustrated in Figure 1, which summarizes the main stages of the structured synthesis adopted in this paper. The literature corpus was assembled from recent peer-reviewed journal articles, conference papers, and authoritative technical reports focusing on: (i) privacy and security risks in GenAI and LLM-based systems, (ii) empirical demonstrations of privacy leakage and adversarial attacks, and (iii) governance, privacy risk assessment, and AI risk management frameworks applicable to organizational settings. Particular attention was given to studies published between 2023 and 2025, reflecting the rapid evolution of GenAI technologies and associated risk landscapes.

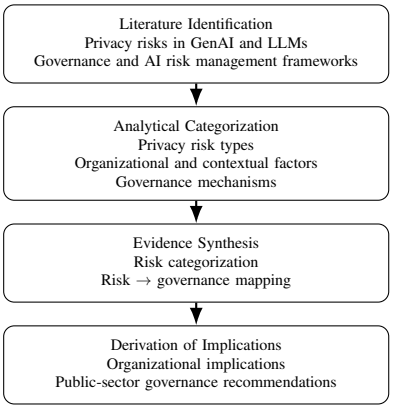

Figure 1: Research approach adopted for the synthesis of privacy risks and governance mechanisms in GenAI.

The analysis followed an iterative qualitative synthesis process. In the first stage, relevant studies were screened and examined to extract recurrent privacy risk patterns, such as data memorization, inference-based leakage, prompt-driven information disclosure, and lifecycle emergent risks. In the second stage, organizational and governance-related mechanisms discussed in the literature including privacy impact assessments, lifecycle-based risk management, accountability structures, and regulatory alignment were identified and grouped thematically. This analytical strategy is consistent with prior guidance on thematic synthesis and mapping studies in software and information systems research (Petersen et al., 2008).

In the third stage, the extracted evidence was consolidated into two complementary synthesis artifacts. The first artifact summarizes representative studies and the privacy risks they identify, highlighting how such risks materialize in concrete organizational and deployment contexts. The second artifact maps recurrent privacy risk categories to governance-oriented mitigation mechanisms, emphasizing the role of organizational structures, processes, and accountability arrangements in complementing technical safeguards. Finally, based on this synthesis, we derived governance recommendations with particular focus on public-sector GenAI deployments, where privacy risks are amplified by legal obligations, societal impact, and transparency requirements.

4 PRIVACY RISKS OF GENAI IN ORGANIZATIONS

The adoption of Generative Artificial Intelligence (GenAI) in organizational environments introduces a distinct and evolving set of privacy risks that differ substantially from those observed in traditional data-driven systems. Unlike conventional machine learning applications, GenAI systems particularly Large Language Models (LLMs) operate through open-ended interaction, probabilistic reasoning, and contextual inference, which fundamentally alters system behavior and risk exposure (Chen et al., 2025; Ruhländer et al., 2024). These characteristics blur system boundaries and complicate the identification, assessment, and mitigation of privacy risks. Consequently, privacy risks associated with GenAI should be understood as socio-technical risks that emerge from the interaction between models, data, users, and organizational practices rather than as isolated technical vulnerabilities, reinforcing the need for governance-oriented risk management approaches (Sattlegger and Bharosa, 2024; Esposito and Tse, 2024).

Data Memorization and Inference-Based Privacy Leakage One of the most widely recognized privacy risks in LLMs is the unintended memorization of personal or sensitive data during training. Empirical studies demonstrate that LLMs may reproduce fragments of training data or enable the inference of sensitive attributes under specific conditions (Ruhländer

et al., 2024; Chen et al., 2025). In organizational contexts, this risk is particularly critical when training or fine-tuning data include customer records, employee information, or regulated datasets.

Beyond direct memorization, recent research shows that privacy leakage may occur through inference-based attacks, such as membership inference and attribute inference, even when models are pre-trained and deployed as black-box services (Chang et al., 2024). These attacks exploit subtle variations in model outputs and are highly context-dependent, meaning that privacy risks may intensify depending on how GenAI systems are embedded into organizational workflows. Importantly, these findings indicate that avoiding sensitive data during training is not sufficient to ensure privacy protection in real-world deployments (Ruhländer et al., 2024).

Prompt-Based Interaction and Information Exfiltration The prompt-based interaction paradigm introduced by GenAI constitutes a major source of privacy risk in organizational settings. Prompts frequently contain contextual information, operational details, or personal data that may be processed, logged, or inadvertently retained by the system. Recent work has demonstrated that prompt injection attacks both direct and indirect can manipulate model behavior to exfiltrate sensitive information (Derner et al., 2024; Chernyshev et al., 2024).

Empirical evidence further shows that personal information can be actively extracted from deployed LLMs via prompt injection techniques, even in the absence of explicit vulnerabilities or privileged access (Schwartzman, 2024). Forensic analyses reveal that such attacks may leverage legitimate system functionalities, complicating detection, attribution, and incident response in organizational environments (Chernyshev et al., 2024). As a result, prompts themselves become organizational artifacts that require explicit governance, oversight, and usage policies.

Lifecycle and Contextual Emergence of Privacy Risks Privacy risks in GenAI systems do not remain static throughout the system lifecycle. Instead, they may emerge or evolve during deployment and use, influenced by contextual factors such as system configuration, access privileges, logging practices, and integration with external tools or data sources. Surveys on privacy and security risks in LLMs highlight that leakage may occur long after deployment, driven by changes in usage patterns or adversarial exploration of model behavior (Chen et al., 2025; Ahi and Valizadeh, 2025).

Several recent studies emphasize that privacy risks in GenAI are strongly dependent on deployment context and usage scenarios, reinforcing the dual-use nature of these technologies (Barberá, 2025; Ahi and Valizadeh, 2025). In practice, organizations often lack visibility into how GenAI systems are used across departments, giving rise to phenomena such as shadow AI, undocumented data flows, and informal reuse of prompts or outputs. These practices amplify privacy risks and undermine accountability, particularly in regulated environments subject to data protection laws such as GDPR (Parliament and Council, 2018) and LGPD (da República, 2018).

Organizational and Governance Related Privacy Risks In addition to technical and interaction level risks, GenAI introduces privacy risks rooted in organizational structures and governance arrangements. Studies on privacy impact assessment and risk management reveal that many organizations rely on fragmented or compliance-oriented practices that fail to capture the evolving risk profile of AI systems (Wairimu et al., 2024). This limitation is especially problematic for GenAI, where privacy risks often arise from collective use, organizational routines, and decision-making processes rather than from individual system components.

Empirical evidence further indicates that perceived privacy risks do not necessarily translate into reduced use of AI systems, as efficiency gains, organizational pressure, and trust in automation frequently outweigh privacy concerns in practice (Otoo et al., 2025). Moreover, the absence of clearly defined roles, responsibilities, and escalation mechanisms complicates incident response and accountability. Recent work on ethical AI and algorithmic governance emphasizes that privacy risks must be embedded into organizational risk management practices rather than treated solely as external compliance requirements (Sattlegger and Bharosa, 2024; Esposito and Tse, 2024).

Limitations of Purely Technical Mitigation Strategies Although numerous technical mitigation strategies have been proposed such as differential privacy, prompt filtering, access control, and output monitoring the literature consistently indicates that these measures are insufficient when applied in isolation. Surveys and empirical studies highlight that technical controls may reduce specific attack vectors but cannot fully address risks arising from contextual inference, evolving usage patterns, or organizational misuse (Chen et al., 2025; Rathod et al., 2025).

Furthermore, privacy risks have been observed

even in systems that adhere to best-practice technical safeguards, underscoring the need for complementary governance-oriented mechanisms (Dotan et al., 2024). These findings reinforce the argument that effective mitigation of privacy risks in GenAI requires a shift from a purely technical perspective toward integrated governance and risk management approaches, which are discussed in the next section.

Table 1 summarizes representative studies and the corresponding privacy risks they identify. Rather than presenting an exhaustive catalog of threats, the table highlights recurring risk patterns observed across empirical studies, surveys, and governance-oriented analyses, including data memorization, inference-based leakage, prompt injection, and organizational governance gaps (Chen et al., 2025; Ruhländer et al., 2024; Derner et al., 2024). The inclusion of application scenarios, when reported by the original studies, further illustrates how privacy risks materialize in concrete organizational contexts, reinforcing the view that such risks are not merely technical artifacts but socio-technical phenomena shaped by system design, usage patterns, and governance arrangements (Wairimu et al., 2024; Sattlegger and Bharosa, 2024).

5 GOVERNANCE MODELS FOR RISK MITIGATION

The dynamic and context-dependent nature of privacy risks in GenAI systems requires mitigation strategies that extend beyond isolated technical controls (Chen et al., 2025; Ruhländer et al., 2024; Chang et al., 2024). As discussed in the previous section, privacy risks in organizational deployments of GenAI emerge from the interaction between models, data, users, and institutional practices, reflecting their inherently socio-technical character (Wairimu et al., 2024; Sattlegger and Bharosa, 2024). Consequently, effective mitigation depends on governance models that integrate technical safeguards with organizational structures, risk management processes, and accountability mechanisms, as emphasized in recent AI risk management and algorithmic governance frameworks (of Standards and Technology, 2023; Dotan et al., 2024; Esposito and Tse, 2024).

From Technical Controls to Governance-Oriented Mitigation A substantial portion of the literature on privacy in Large Language Models focuses on technical mitigation strategies, such as differential privacy, prompt filtering, access control, and output monitoring. While these mechanisms are necessary to reduce specific attack vectors, empirical studies consistently show that they are insufficient when applied in isolation, particularly in complex organizational environments (Chen et al., 2025; Rathod et al., 2025). Technical controls often assume stable system boundaries and predictable data flows, assumptions that do not hold for GenAI systems characterized by open-ended interaction and evolving usage patterns.

Governance-oriented mitigation approaches address this limitation by framing privacy risks as organizational risks that require coordinated decision-making, oversight, and continuous adaptation. Rather than seeking to eliminate all technical vulnerabilities, governance models aim to manage acceptable levels of risk through policies, roles, procedures, and monitoring mechanisms that evolve alongside the system (Sattlegger and Bharosa, 2024).

Risk Management Frameworks for Generative AI Risk management frameworks provide a foundational structure for governing privacy risks in GenAI systems. The NIST Artificial Intelligence Risk Management Framework (AI RMF 1.0) proposes a lifecycle based approach to identifying, assessing, and managing AI risks, emphasizing governance, accountability, and continuous monitoring (of Standards and Technology, 2023). Although the AI RMF is technology agnostic, its core functions: *Govern*, *Map*, *Measure*, and *Manage* offer a useful lens for structuring privacy risk mitigation in organizational GenAI deployments.

Recent work has proposed maturity models to operationalize the AI RMF in practice, highlighting that organizations vary significantly in their ability to systematically manage AI risks (Dotan et al., 2024). These models emphasize the need for clearly defined roles, documented processes, and performance indicators to support consistent decision-making. In the context of GenAI, maturity based approaches are particularly relevant, as they enable organizations to incrementally improve governance practices while accommodating rapid technological change.

Privacy Impact Assessment and Continuous Risk Evaluation Privacy Impact Assessment (PIA) and Privacy Risk Assessment (PRA) are widely recognized instruments for supporting compliance with data protection regulations. However, systematic reviews indicate that existing PIA and PRA methodologies are often applied as one-off compliance exercises and lack empirical validation in adaptive AI contexts (Wairimu et al., 2024). For GenAI systems, where privacy risks may emerge during use rather than at design time, such static assessments are insufficient.

Governance models for GenAI therefore require a shift toward continuous and iterative risk evaluation

Table 1: Summary of privacy risks identified in Generative AI studies

Study	Privacy Risks Identified	Application Scenario
(Chen et al., 2025)	Data memorization, inference-based leakage, training data exposure	General purpose LLMs
(Chang et al., 2024)	Context-aware membership inference attacks	Pre-trained LLMs in black-box settings
(Derner et al., 2024)	Prompt injection, unauthorized information disclosure	Prompt-based interaction systems
(Schwartzman, 2024)	Exfiltration of personal information via prompt manipulation	Deployed conversational AI systems
(Chernyshev et al., 2024)	Indirect prompt injection and covert data exfiltration	LLM agents and tool-augmented systems
(Wairimu et al., 2024)	Inadequacy of traditional PIA/PRA for adaptive AI	Organizational risk assessment practices
(Sattlegger and Bharosa, 2024)	Governance gaps, lack of accountability	Public sector AI systems
(Esposito and Tse, 2024)	Insufficient algorithmic governance mechanisms	Governmental GenAI deployments

practices. This includes periodic reassessment of privacy risks, monitoring of system behavior and usage patterns, and mechanisms for updating controls in response to emerging threats (Dotan et al., 2024). Embedding PIA and PRA within broader organizational risk management processes helps ensure that privacy considerations remain visible and actionable throughout the system lifecycle.

Algorithmic Governance and Organizational Accountability Beyond formal risk management frameworks, recent research emphasizes the role of algorithmic governance in mitigating privacy risks associated with GenAI. Algorithmic governance approaches advocate for embedding ethical, legal, and privacy considerations into organizational decision-making structures, rather than treating them as external compliance requirements (Esposito and Tse, 2024). This includes the establishment of cross-functional governance bodies, clear allocation of responsibilities, and escalation procedures for privacy related incidents.

Studies in public-sector contexts demonstrate that the absence of explicit governance arrangements often leads to fragmented accountability and reactive risk management practices (Sattlegger and Bharosa, 2024). Although these findings originate in governmental settings, they are equally applicable to private organizations deploying GenAI, where decentralized adoption and shadow AI practices can undermine privacy protection efforts. Algorithmic governance thus provides a mechanism to align technical development, operational use, and regulatory compliance under a unified governance framework.

Regulatory Alignment and Organizational Governance Regulatory frameworks such as the GDPR(Parliament and Council, 2018) and LGPD(da República, 2018) establish legal obligations for organizations processing personal data, including requirements for transparency, accountability, and risk-based decision-making. In parallel, emerging AI specific regulations, such as the European Union Artificial Intelligence Act, reinforce the need to integrate privacy protection into broader AI governance structures (Sonsini and Parliament, 2024). Comparative studies of data protection laws and privacy frameworks indicate a growing convergence toward risk based and governance-oriented approaches (Rocha and Canedo, 2025; Ferrão et al., 2024).

Governance models for GenAI must therefore support regulatory alignment while remaining flexible enough to address evolving technical risks. This includes maintaining documentation, audit trails, and decision rationales that demonstrate compliance, as well as ensuring that governance practices are proportionate to the risk level and application context of the GenAI system.

Toward Integrated Governance Models for GenAI Privacy The literature suggests that effective mitigation of privacy risks in GenAI requires integrated governance models that combine technical controls, risk management frameworks, and organizational accountability mechanisms. Rather than prescribing a single prescriptive solution, existing studies point to the importance of tailoring governance practices to organizational context, maturity level, and regulatory environment (Dotan et al., 2024). By positioning privacy risk mitigation as an ongoing governance activity rather than a static technical task, organizations can better manage the uncertainties and emergent behaviors associated with Generative AI systems. This governance-oriented perspective provides a foundation for addressing the complex privacy challenges identified in this study and directly responds to the research question guiding this paper.

To operationalize the relationship between privacy risks and governance-oriented mitigation strategies, Table 2 synthesizes evidence from the literature by mapping recurrent privacy risk categories in Generative AI to corresponding governance mechanisms. Rather than proposing ad hoc technical countermeasures, the table consolidates how existing studies and regulatory frameworks emphasize organizational controls, risk management processes, and accountability structures as important complements to technical safeguards (Chen et al., 2025; Sattlegger and Bharosa, 2024; of Standards and Technology, 2023). By grounding governance mechanisms in empirical

risk evidence and regulatory instruments, this synthesis highlights how privacy risk mitigation in GenAI systems requires an integrated, socio-technical governance approach aligned with both organizational practice and data protection obligations (Parliament and Council, 2018; da República, 2018; Rocha and Canedo, 2025).

6 DISCUSSION

This study investigated the main privacy risks associated with the use of GenAI in organizational contexts and examined how governance models can mitigate such risks. The findings indicate that privacy risks in GenAI systems are not limited to isolated technical vulnerabilities but instead emerge as dynamic and socio-technical phenomena shaped by interaction patterns, organizational practices, and governance arrangements. This observation has important implications for both research and practice.

Implications for Research. From a research perspective, the synthesis of recent studies reveals a shift in how privacy risks in AI systems should be conceptualized. While prior work on privacy-preserving machine learning has largely focused on static data pipelines and technical safeguards, the literature analyzed in this paper highlights the limitations of this approach for GenAI systems (Chen et al., 2025; Ruhländer et al., 2024). The open-ended interaction paradigm of LLMs, combined with probabilistic reasoning and contextual inference, challenges traditional assumptions about system boundaries and risk controllability.

The mapping between privacy risks and governance mechanisms further suggests that effective mitigation requires interdisciplinary perspectives that integrate software engineering, risk management, organizational studies, and regulatory analysis. In this sense, this paper contributes to ongoing discussions on socio-technical risk governance by providing a structured synthesis that explicitly links empirical evidence of privacy risks to governance-oriented mitigation strategies (Sattlegger and Bharosa, 2024). Future research may build upon this mapping to empirically evaluate the effectiveness of specific governance mechanisms in different organizational settings.

Implications for Organizational Practice. For practitioners, the findings underscore the inadequacy of relying solely on technical controls to manage privacy risks in GenAI deployments. Although mechanisms such as access control, prompt filtering, and monitoring are necessary, they do not address risks arising from organizational factors such as informal adoption, unclear accountability, or misaligned incentives (Wairimu et al., 2024). The persistence of GenAI use despite perceived privacy risks, as observed in empirical studies, further reinforces the need for governance models that shape organizational behavior rather than depending exclusively on individual awareness or compliance (Otoo et al., 2025).

The proposed risk to governance mapping illustrates how organizations can operationalize privacy protection through structured governance practices, including lifecycle-based risk management, role definition, and continuous assessment aligned with established frameworks such as the NIST AI Risk Management Framework (of Standards and Technology, 2023). These practices enable organizations to adapt to evolving risk profiles while maintaining accountability and regulatory compliance.

Regulatory and Policy Implications. The discussion also highlights important implications for regulators and policymakers. Data protection regulations such as the GDPR and LGPD already emphasize risk-based and accountability-driven approaches to privacy protection (Parliament and Council, 2018; da República, 2018). However, the emergence of GenAI systems exposes gaps between regulatory intent and organizational practice, particularly when governance structures are fragmented or compliance-oriented.

Recent AI specific regulatory initiatives, including the European Union Artificial Intelligence Act, further reinforce the need to integrate privacy risk management into broader AI governance frameworks (Sonsini and Parliament, 2024; Kim et al., 2025). The findings of this study suggest that regulatory effectiveness depends not only on legal requirements but also on the availability of practical governance models that organizations can adopt and adapt. In this context, governance frameworks and maturity models play a critical role in translating regulatory principles into actionable organizational practices (Dotan et al., 2024).

7 RECOMMENDATIONS FOR PUBLIC SECTOR GOVERNANCE

The public sector represents a particularly sensitive and high-risk context for the deployment of GenAI, as AI-supported decisions may directly affect fun-

Table 2: Mapping privacy risks of Generative AI to governance mitigation mechanisms

Privacy Risk Category	Representative Evidence	Governance-Oriented Mitigation Mechanisms
Data memorization and inference-based leakage	Surveys and empirical attacks on LLMs demonstrating memorization and membership inference (Chen et al., 2025; Chang et al., 2024; Ruhländer et al., 2024)	Data governance policies, restrictions on training and fine-tuning datasets, continuous privacy risk assessment, and accountability aligned with data protection regulations (e.g., GDPR, LGPD) (Parliament and Council, 2018; da República, 2018; Rocha and Canedo, 2025)
Prompt-based information disclosure and exfiltration	Prompt injection taxonomies and forensic analyses showing extraction of sensitive data (Derner et al., 2024; Chernyshev et al., 2024; Schwartzman, 2024)	Prompt governance policies, role-based access control, logging and auditing of interactions, incident response procedures, and user accountability mechanisms (of Standards and Technology, 2023; Esposito and Tse, 2024)
Context-dependent and lifecycle emergent privacy risks	Evidence that privacy risks evolve during deployment and use, depending on context and organizational practices (Ahi and Valizadeh, 2025; Barberá, 2025)	Lifecycle based AI risk management, continuous monitoring, and maturity based governance approaches grounded in the NIST AI RMF (of Standards and Technology, 2023; Dotan et al., 2024)
Fragmented organizational accountability and shadow AI	Studies highlighting governance gaps, informal AI adoption, and weak accountability structures (Sattlegger and Bharosa, 2024; Wairimu et al., 2024)	Formal governance bodies, clear role definition, integration of privacy risk into enterprise risk management, and cross-functional oversight (for Economic Co-operation and (OECD), 2024a; for Economic Co-operation and (OECD), 2024b)
Misalignment between technical controls and regulatory obligations	Empirical evidence of compliance driven but ineffective privacy practices (Wairimu et al., 2024; Matos et al., 2025)	Regulatory alignment through DPIA/PIA integration, documentation, auditability, and governance structures aligned with GDPR, LGPD, and emerging AI regulation (Parliament and Council, 2018; da República, 2018; Sonsini and Parliament, 2024; Kim et al., 2025)

damental rights, access to public services, and trust in government institutions. Unlike private organizations, public administrations operate under strict legal mandates, transparency obligations, and accountability requirements, which amplify the impact of privacy risks associated with GenAI systems. Based on the risks and governance mechanisms discussed in this paper, this section presents a set of recommendations aimed at mitigating privacy risks in public-sector GenAI deployments.

Institutional Governance and Accountability Structures A recurring finding in the literature is that privacy risks in public-sector AI systems are exacerbated by fragmented governance structures and unclear allocation of responsibilities (Sattlegger and Bharosa, 2024). To address this challenge, public organizations should establish formal institutional governance arrangements for GenAI, including cross-functional committees or steering groups responsible for overseeing AI use across departments. These bodies should integrate legal, technical, and policy expertise to ensure that privacy considerations are consistently embedded into decision-making processes.

Clear accountability mechanisms are necessary to prevent diffusion of responsibility in complex GenAI deployments. Governance frameworks should explicitly define roles such as data controllers, AI system owners, risk managers, and oversight authorities, ensuring alignment with data protection regulations such as the LGPD and GDPR (da República, 2018; Parliament and Council, 2018). Recent studies on algorithmic governance in government contexts emphasize that embedding accountability into organizational structures is more effective than relying solely on high-level ethical principles (Esposito and Tse, 2024).

Risk-Based Deployment and Continuous Privacy Assessment Public-sector organizations should adopt a risk-based approach to GenAI deployment, recognizing that privacy risks vary according to application domain, data sensitivity, and societal impact. Privacy Impact Assessments (PIAs) and Privacy Risk Assessments (PRAs) should be treated as living instruments rather than one-off compliance artifacts, particularly given the adaptive and evolving behavior of GenAI systems (Wairimu et al., 2024).

In line with the NIST AI Risk Management Framework, public administrations are encouraged to implement continuous monitoring and reassessment practices throughout the AI system lifecycle (of Standards and Technology, 2023). This includes periodic review of prompts, outputs, and usage patterns, as well as mechanisms for detecting emergent risks such as prompt-based information leakage or inference attacks (Chen et al., 2025). Maturity models grounded in AI risk management have been proposed as a practical means to support incremental improvement of governance capabilities in public institutions (Dotan et al., 2024).

Data Governance and Inter-Organizational Coordination Effective mitigation of privacy risks in public-sector GenAI deployments depends on robust data governance practices. Public administrations often operate within complex data ecosystems involving multiple agencies, external service providers, and shared digital platforms. Studies on data governance highlight that weak coordination and unclear data stewardship roles significantly increase privacy risks in AI-enabled public services (for Economic Co-operation and (OECD), 2024a; for Economic Co-operation and (OECD), 2024b).

To mitigate these risks, public-sector organizations should establish explicit policies governing the use of personal and sensitive data in GenAI systems, including restrictions on training, fine-tuning,

and prompt usage. Inter-organizational agreements should clarify responsibilities for data protection, auditing, and incident response, particularly when GenAI services are procured from external vendors or deployed as cloud-based solutions (Gudepu et al., 2025).

Capacity Building and Organizational Maturity Several studies emphasize that technical safeguards alone are insufficient to manage privacy risks if organizational actors lack the competencies required to understand and apply governance mechanisms effectively (Matos et al., 2025; Esposito and Tse, 2024; Falcão and Canedo, 2024). Capacity building should therefore be treated as a core component of public-sector GenAI governance. This includes targeted training for policymakers, managers, and technical staff on privacy risks, responsible AI practices, and regulatory obligations(Spósito et al., 2025a; Martins et al., 2025; Gonçalves et al., 2024).

Organizational maturity models provide a useful framework for guiding capacity building efforts, enabling public institutions to assess their current governance capabilities and identify priority areas for improvement (Dotan et al., 2024; Saraiva et al., 2025). By progressively enhancing governance maturity, public-sector organizations can better align GenAI innovation with privacy protection, accountability, and public trust.

Regulatory Alignment and Transparency Public-sector GenAI governance should explicitly align with existing and emerging regulatory frameworks. The LGPD, GDPR, and the EU Artificial Intelligence Act collectively emphasize principles of transparency, accountability, and risk-based governance (da República, 2018; Parliament and Council, 2018; Sonsini and Parliament, 2024). Public administrations should operationalize these principles through documentation, audit trails, and explainability mechanisms that support both internal oversight and external scrutiny.

Transparency toward citizens is particularly critical in the public sector. Governance models should ensure that the use of GenAI systems is clearly communicated, that individuals are informed about how their data are processed, and that accessible mechanisms for redress and contestation are available (Sattlegger and Bharosa, 2024). Such measures are necessary to maintaining public trust and legitimacy in the use of GenAI within governmental institutions.

Key Findings Summary

The findings indicate that privacy risks associated with Generative Artificial Intelligence (GenAI) in organizational contexts are dynamic, context-dependent, and predominantly socio-technical in nature. Rather than arising solely from isolated technical vulnerabilities, these risks emerge from the interaction between models, data, users, and organizational practices. Empirical evidence shows that technical safeguards such as access control, prompt filtering, and anonymization are necessary but insufficient to address risks such as data memorization, inference-based leakage, and prompt-based information exfiltration. Effective mitigation therefore depends on governance models that integrate technical controls with organizational structures, continuous risk management processes, and clear accountability mechanisms. In particular, risk-based governance frameworks, lifecycle oriented assessment, and institutional accountability are essential to managing privacy risks in GenAI deployments, especially in highly regulated and high impact contexts such as the public sector.

7.1 Threats to Validity

This study is subject to several limitations. First, the analysis is based on a synthesis of recent literature rather than on primary empirical data, which may reflect biases or gaps present in the selected studies. Although care was taken to include surveys, empirical analyses, and governance-oriented research from multiple venues, relevant works may have been inadvertently omitted.

Second, the mapping between privacy risks and governance mechanisms is interpretative in nature and reflects the authors' analytical perspective. While this synthesis is grounded in established frameworks and empirical evidence, alternative categorizations or governance interpretations may lead to different mappings. Finally, the focus on organizational and public-sector contexts may limit the generalizability of the findings to other domains or highly experimental GenAI applications. Future empirical studies are needed to validate and refine the proposed risk governance relationships in real-world deployments.

8 CONCLUSION

This study examined the main privacy risks associated with the adoption of Generative Artificial Intelligence (GenAI) in organizational contexts and investigated how governance models can mitigate these risks. The analysis of recent literature indicates that privacy risks in GenAI systems are dynamic, context-dependent, and predominantly socio-technical, emerging from the interaction between models, data, users, and organizational practices rather than from isolated technical flaws.

The findings demonstrate that purely technical mitigation strategies, while necessary, are insufficient to address risks such as data memorization, inference-based leakage, and prompt-driven information disclosure. Effective mitigation requires integrated governance models that combine technical safeguards with organizational structures, continuous risk management processes, and clearly defined accountability mechanisms. Frameworks such as risk-based governance, lifecycle oriented assessment, and algorithmic governance provide a structured foundation for managing privacy risks in GenAI deployments. In particular, the study highlights the relevance of governance-oriented approaches for the public sector, where GenAI systems operate under heightened legal, ethical, and societal constraints. The proposed mapping between privacy risks and governance mechanisms contributes to bridging the gap between abstract regulatory principles and actionable organizational practices.

As future work, empirical studies are needed to evaluate the effectiveness of specific governance mechanisms in real world GenAI deployments and to refine maturity models tailored to different organizational and regulatory contexts. Further research may also explore sector specific adaptations of governance frameworks and the long-term impact of GenAI governance practices on public trust and institutional accountability.

ACKNOWLEDGEMENTS

This study was financed in part by the Project No. TED 34/2024 “Pesquisa Aplicada em Privacidade e Segurança da Informação na Diretoria de Privacidade e Segurança da Informação da Secretaria de Governo Digital” – Diretoria de Privacidade e Segurança da Informação (DPSI)/Centro de Excelência em Privacidade e Segurança (CEPS)/Secretaria de Governo Digital (SGD) do Ministério da Gestão e da Inovação (MGI) em Serviços Públicos; and the National Council for Scientific and Technological Development (CNPq) under Grants No. 300883/2025-0 and No. 406266/2025-5.

REFERENCES

Ahi, K. and Valizadeh, S. (2025). Large language models (llms) and generative ai in cybersecurity and privacy: A survey of dual-use risks, ai-generated malware, explainability, and defensive strategies. In *2025 Silicon Valley Cybersecurity Conference (SVCC)*, pages 1–8.

Azevedo, L. F. D. and Canedo, E. D. (2025). A structured checklist approach to evaluating transparency and privacy in brazilian digital services. In *Proceedings of the 24th Brazilian Symposium on Software Quality, SBQS 2025, São José dos Campos, SP, Brazil, November 4-7, 2025*, pages 400–410. SBC.

Barberá, I. (2025). Ai privacy risks & mitigations—large language models (llms). *European Data Protection Board. Available online: https://www. edpb. europa. eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms. pdf (accessed on 12 June 2025)*.

Chang, H., Shamsabadi, A. S., Katevas, K., Haddadi, H., and Shokri, R. (2024). Context-aware membership inference attacks against pre-trained large language models. *CoRR*, abs/2409.13745.

Chen, K., Zhou, X., Lin, Y., Feng, S., Shen, L., and Wu, P. (2025). A survey on privacy risks and protection in large language models. *J. King Saud Univ. Comput. Inf. Sci.*, 37(7).

Chernyshev, M., Baig, Z. A., and Doss, R. (2024). [short paper] forensic analysis of indirect prompt injection attacks on LLM agents. In *5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications, TPS-ISA, Atlanta, GA, USA, November 1-4,*, pages 409–411. IEEE.

da República, Presidência, N. C. (2018). Brazilian general data protection law (lgpd). *Nartional Congress, accessed in April 10, 2022*, 1(1):1–31.

Derner, E., Batistic, K., Zahálka, J., and Babuska, R. (2024). A security risk taxonomy for prompt-based interaction with large language models. *IEEE Access*, 12:126176–126187.

Dotan, R., Blili-Hamelin, B., Madhavan, R., Matthews, J., and Scarpino, J. (2024). Evolving AI risk management: A maturity model based on the NIST AI risk management framework. *CoRR*, abs/2401.15229.

Esposito, M. and Tse, T. (2024). Mitigating the risks of generative AI in government through algorithmic governance. In *Proceedings of the 25th Annual International Conference on Digital Government Research, DGO 2024, Taipei, Taiwan, June 11-14, 2024*, pages 605–609. ACM.

Falcão, F. D. S. and Canedo, E. D. (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, pages 373–382. ACM.

Ferrão, S. É. R., Silva, G. R. S., Canedo, E. D., and Mendes, F. F. (2024). Towards a taxonomy of privacy requirements based on the LGPD and ISO/IEC 29100. *Inf. Softw. Technol.*, 168:107396.

for Economic Co-operation, O. and (OECD), D. (2024a). Data governance and privacy: Synergies and areas of international co-operation. *OECD: Paris, France*.

for Economic Co-operation, O. and (OECD), D. (2024b). Oecd privacy guidelines. *OECD: Paris, France*, pages 1–19.

Gonçalves, C. D., de Paoli Menescal, E., de Mendonça, F. L. L., and Canedo, E. D. (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*, pages 147–157. ACM.

Gudepu, B. K., Pemmasani, P. K., Gonugunta, K. C., and Jegajothi, B. (2025). Ai-augmented data privacy risk assessment framework for smart governance applications. In *2025 3rd International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, pages 275–283.

Kim, B., Jeong, S., Cho, B., and Chung, J. (2025). AI governance in the context of the EU AI act. *IEEE Access*, 13:144126–144142.

Martins, M. E., Aguiar, Y. P. C., and Saraiva, J. (2025). Assessment of competences for LGPD DPO through ANPD standard and information systems curriculum. In *Proceedings of the 21st Brazilian Symposium on Information Systems, SBSI 2025, Recife, Brazil, May 19-23, 2025*, pages 565–574. SBC.

Matos, A., Patrício, M., Nicolau, M. I., Canedo, E. D., Pereira, J. A., and Uchôa, A. G. (2025). Data privacy in software practice: Brazilian developers' perspectives. *J. Internet Serv. Appl.*, 16(1):299–319.

OAIC, A. G. (1988). The Privacy Act.

of California Department of Justice, S. (2018). California consumer privacy act.

of Standards, N. I. and Technology (2023). Artificial intelligence risk management framework (ai rmf 1.0). *NIST AI 100-1*, pages 1–48.

of the Whole House on the State of the Union, C. (2022). American Data Privacy and Protection Act (ADPPA).

Otoo, B. A. A., Osei-Frimpong, K., and Islam, N. (2025). Do perceived privacy risks of AI matter? A longitudinal study on the drivers of continued use of intelligent voice assistants. *IEEE Trans. Engineering Management*, 72:2521–2534.

Parliament, T. E. and Council, T. (2018). General Data Protection Regulation (GDPR): EU Data Protection Rules.

Pautasso, M. (2019). The structure and conduct of a narrative literature review. *A guide to the scientific career: Virtues, communication, research and academic writing*, pages 299–310.

Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. (2008). Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering, EASE 2008, University of Bari, Italy, 26-27 June 2008*, Workshops in Computing. BCS.

Rathod, V., Nabavirazavi, S., Zad, S., and Iyengar, S. S. (2025). Privacy and security challenges in large language models. In *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 00746–00752.

Rocha, L. D. and Canedo, E. D. (2025). Optimizing compliance: Comparative study of data laws and privacy frameworks. *J. Internet Serv. Appl.*, 16(1):431–452.

Ruhländer, L., Popp, E., Stylidou, M., Khan, S., and Svetinovic, D. (2024). On the security and privacy implications of large language models: In-depth threat analysis. In *2024 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics, Copenhagen, Denmark, August 19-22, 2024*, pages 543–550. IEEE.

Saraiva, J., de Souza, C., and Soares, S. (2025). MMAI-LGPD: A maturity model for governance and data compliance in information systems institutions. In *Proceedings of the 21st Brazilian Symposium on Information Systems, SBSI 2025, Recife, Brazil, May 19-23, 2025*, pages 788–797. SBC.

Sattlegger, A. and Bharosa, N. (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*, pages 70–80. ACM.

Schwartzman, G. (2024). Exfiltration of personal information from chatgpt via prompt injection. *CoRR*, abs/2406.00199.

Sonsini, W. and Parliament, T. E. (2024). The eu artificial intelligence act. *European Union*, pages 1–10.

Spósito, S. L., Alves, K., Nunes, R. R., Ferreira, L. R., and Canedo, E. D. (2025a). Structuring privacy and information security competencies for public sector roles: A framework for enhancing software quality and LGPD compliance. In *Proceedings of the 24th Brazilian Symposium on Software Quality, SBQS 2025, São José dos Campos, SP, Brazil, November 4-7, 2025*, pages 365–375. SBC.

Spósito, S. L., Targino, J. F. G., Silva, G. R. S., de Melo, L. P., de Paula Porto, D., de Mendonça, F. L. L., and Canedo, E. D. (2025b). A comprehensive review of techniques, methods, processes, frameworks, and tools for privacy requirements. *J. Internet Serv. Appl.*, 16(1):508–529.

Wairimu, S., Iwaya, L. H., Fritsch, L., and Lindskog, S. (2024). On the evaluation of privacy impact assessment and privacy risk assessment methodologies: A systematic literature review. *IEEE Access*, 12:19625–19650.

Wang, P. and Zare, H. (2023). A case study of privacy protection challenges and risks in ai-enabled healthcare app. In *IEEE Conference on Artificial Intelligence,*

CAI 2023, Santa Clara, CA, USA, June 5-6, 2023, pages 296–297. IEEE.