International Journal of Frontiers in
Science and Technology Research
Journal homepage: https://frontiersrj.com/journals/ijfstr/
ISSN: 2783-0446 (Online)

(RESEARCH ARTICLE)

Check for updates

SECURITY AND PRIVACY CHALLENGES OF RAG SYSTEMS

Firoz Mohammed Ozman *

Solutions Architect, Enterprise Architecture, Anecca Ideas Corp, Toronto, Canada.
* Corresponding Author

ORCID Details
Firoz Mohammed Ozman: https://orcid.org/0009-0006-2381-7162

International Journal of Frontiers in Science and Technology Research, 2026, 10(02), 001–013

Publication history: Received on 07 July 2026; revised on 13 August 2026; accepted on 15 August 2026

Article DOI: https://doi.org/10.53294/ijfstr.2026.10.2.0022

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Retrieval-Augmented Generation (RAG) systems enable robust knowledge integration for large language models but also pose significant security and privacy risks. RAG systems combine two components: a retriever, which searches external data sources for relevant information, and a generator, typically a large language model that uses both the retrieved documents and user queries to produce answers. This study conducts a systematic literature review to assess these challenges using Socio-Technical Systems Theory (which considers interactions among people, technology, and organizational context) and Privacy by Design (PbD, a framework for embedding privacy into system design). Addressing five objectives, the research detects and classifies privacy attacks, evaluates risks throughout the storage, retrieval, and generation phases, scrutinizes measurement methods, contrasts mitigation strategies, and introduces a unified solution. The work culminates in the Integrated Privacy-Preserving RAG Framework (IPRAG), a five-tier architecture supported by a three-phase deployment protocol. This study presents a detailed, actionable approach to constructing secure, privacy-focused RAG systems.

Keywords: Retrieval-Augmented Generation, RAG privacy, RAG security, Privacy by Design, differential privacy, socio-technical systems, IPRAG, membership inference, data poisoning

1 INTRODUCTION

Large language models (LLMs) are constrained by static training data that rapidly becomes outdated, leading to potentially inaccurate outputs or hallucinations (instances where the model generates information not supported by its data) and a lack of transparency in the generation process (difficulty understanding how outputs are produced). On top of these limitations, retrieval-augmented generation (RAG)- a technique that supplements LLMs with information fetched from external data sources-introduces additional complexities for academic and research activities. RAG requires effective data management and MLOps (Machine Learning Operations, the practices for deploying and maintaining machine learning models) to synthesize diverse, evolving data sources, potentially using data meshes (a decentralized data architecture strategy). However, this approach risks introducing new biases from retrieved data and reduces the interpretability of large datasets (the ability to understand and analyze the data and system outputs). The system's reliability depends on accurate retrieval methods for document ranking and relevance (Klesel & Wittmann,

2025). Furthermore, when handling sensitive data, RAG systems face substantial privacy threats, as adversaries may exploit tailored prompts to extract confidential documents from external databases (He et al., 2025).

1.1 Problem Statement

Despite the accelerated adoption of RAG systems, the privacy risks inherent in their architecture remain poorly defined and inadequately mitigated. Problems such as unintended data exposure, retrieval poisoning, prompt injection, and inference attacks present concrete threats to individuals and organizations. Existing security frameworks are primarily tailored to static machine learning models and fail to account for the dynamic, distributed configuration of RAG systems. Furthermore, collaboration between attack detection, privacy assessment, and risk-reduction strategies is minimal, resulting in a disjointed approach in contemporary research and implementation.

Aim

To critically evaluate privacy risks in RAG systems and propose mitigation strategies focusing on attacks, measurement techniques, and control mechanisms.

Objectives

- To identify the basic concepts of RAG systems
- To identify and categorize privacy attacks in RAG systems
- To analyze privacy risks across retrieval, generation, and storage stages
- To evaluate existing privacy measurement techniques
- To compare mitigation strategies including technical and governance controls
- To propose an integrated framework for privacy-preserving RAG systems

Research Questions

- What are the major privacy attacks affecting RAG systems?
- How do RAG architectures introduce new forms of privacy risk?
- What methods are used to measure privacy leakage in RAG systems?
- What mitigation strategies are effective in reducing privacy risks?

1.2 Research Rationale

The growing deployment of RAG systems in sensitive sectors such as healthcare, finance, and enterprise knowledge management underscores the urgent need for robust privacy protection (He et al, 2025). As traditional AI security literature does not fully capture the emerging privacy risks posed by RAG's unique architecture, this study addresses these gaps. By providing new insights into these challenges, the findings will help developers, organizations, and policymakers implement safer AI systems that balance performance and privacy.

2 LITERATURE REVIEW

2.1 What is RAG?

Retrieval-augmented generation (RAG) refers to artificial intelligence (AI) systems that enhance the performance of text-generating models by integrating information from external databases. Rather than relying exclusively on static training data, RAG systems access both structured sources, such as spreadsheets, and unstructured sources, such as articles, in real time. This approach improves the accuracy and relevance of generated outputs. By combining retrieval and generation, RAG systems enhance model accuracy, flexibility, and scalability (Panda & Mukherjee, 2025). Large language models (LLMs) are trained on extensive datasets to perform various language-related tasks, including question answering, translation, and sentence completion. RAG further augments these models by incorporating up-to-date or organization-specific resources without requiring model retraining, thereby providing a cost-effective means of maintaining current and precise outputs (Amazon, 2024). The RAG workflow begins by retrieving pertinent material from collections or databases in response to user queries, ensuring the selection of the most relevant and up-to-date information. The text-generating model then processes this content to produce responses. Consequently, RAG systems offer greater precision and adaptability compared to conventional AI models. For instance, in the financial sector, a RAG system can extract data from recent market reports to generate compliance documents that reflect current regulations (Malali, 2025).

2.2 Benefits and importance of RAG

Large language models (LLMs) are subject to limitations such as knowledge gaps, lack of realism, and hallucinations. Retrieval-augmented generation (RAG) addresses these deficiencies by integrating external repositories, including databases, into LLMs. RAG is particularly effective for knowledge-driven dialogues that require frequent updates and outperforms alternative approaches by eliminating the need to retrain LLMs for specific applications (Bazzi & Gaith, 2024). The principal advantage of RAG lies in its ability to generate context-aware responses, in contrast to conventional models that often produce generic outputs. For example, in financial documentation, RAG retrieves up-to-date regulatory guidance to generate compliance reports that meet the latest mandates. It also improves precision by sourcing precise data before generating output, thereby reducing the risk of errors in critical tasks such as tax submissions and regulatory filings. In legal practice, RAG identifies relevant case law and precedents for efficient summarization or report creation, enhancing both efficiency and legal accuracy. In healthcare, RAG utilizes patient records and recent research to provide tailored treatment recommendations for diverse medical histories and queries. Customer service chatbots benefit from accessing knowledge bases and FAQs to deliver specific, reliable responses, thereby increasing response speed and customer satisfaction. Crucially, RAG's adaptability allows for the immediate integration of new data or regulatory changes, enabling organizations to maintain compliance in dynamic environments (Malali, 2025).

2.3 Disadvantages of RAG system

RAG systems present several significant disadvantages. They require advanced infrastructure to manage external knowledge bases, and the accuracy of their outputs depends on the reliability of the sources they retrieve. Additionally, RAG systems may introduce privacy and security vulnerabilities when handling sensitive information, as retrieval mechanisms can inadvertently disclose confidential data or be exploited through prompt injection or data poisoning. Other limitations include increased response latency resulting from real-time retrieval and heightened compliance complexity due to evolving data storage and usage requirements (Panda & Mukherjee, 2025).

2.4 RAG system architecture

RAG's structure follows three steps, beginning with indexing. Here, external documents are split into smaller units, such as paragraphs or sentences. These are turned into embedding vectors-numerical summaries of the text-by models like BERT (a language understanding model), then stored in a specialized vector database, such as FAISS. This setup enables rapid similarity searches across very large datasets. The next step is retrieval, triggered when a user sends a question. The system converts the question into a matching vector using the same model. It searches the vector database using cosine similarity (a measure of vector similarity) to find the most relevant results, typically the top 5 to 10. Retrieval accuracy relies on how well the embedding model represents meaning. Finally, in the generation step, the language model (such as a GPT-based one) combines the question and the retrieved data to produce a thoughtful answer, mixing its internal knowledge with new, up-to-date material. This retrieve-and-read method is more accurate and traceable than relying solely on LLMs, and recent systems can incorporate additional sources, such as graphs, to improve reliability (Polan, 2025).

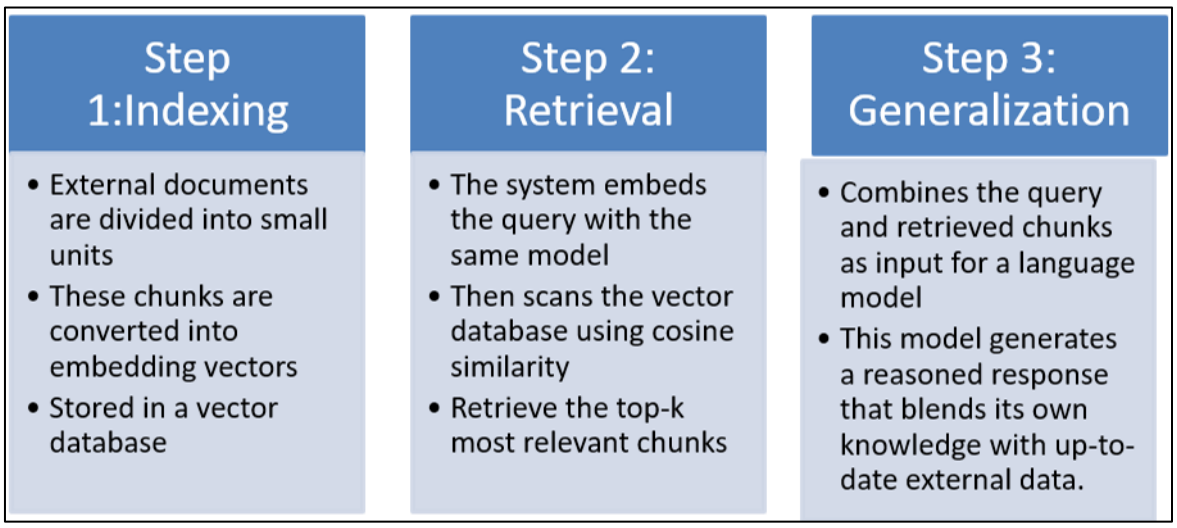

Figure 1: RAG architecture (Polan,2025)

2.5 Implications of privacy risks associated with the RAG system

The privacy risks associated with RAG systems have far-reaching effects, especially under strict data protection laws such as GDPR and CCPA. Organizations that use these systems must protect sensitive data while ensuring the models remain useful and effective. At the same time, ethical issues are growing more important. Concerns include intellectual property, the creation of biased or misleading content, and the wider social effects of powerful AI technologies. For these reasons, creating privacy-preserving methods for RAG is both a technical necessity and an ethical duty (Binwal & Chopra, 2024).

2.6 Theoretical framework

This study is anchored in the integration of Socio-Technical Systems Theory and Privacy by Design (PbD), which together offer a comprehensive framework for analyzing privacy risks in Retrieval-Augmented Generation (RAG) systems. The adoption of advanced technologies in workplaces has significantly transformed human-machine collaboration (Ozman, 2025). Socio-Technical Systems Theory (STST) conceptualizes organizations as combinations of social and technical elements that must function cohesively. The theory emphasizes the interdependence of people, technology, and the environment, positing that optimal performance is achieved when both human and technical factors are designed in tandem. The principle of joint optimization advocates organizing work to enhance both efficiency and employee well-being, in contrast to traditional approaches that prioritize technology and subsequently adapt human roles, which can result in suboptimal outcomes and adverse social effects (Gorejena et al., 2016). In the context of RAG systems, this perspective is particularly relevant, as privacy risks stem not only from technical components, such as large language models or retrieval mechanisms, but also from human behavior, organizational policies, and varied usage scenarios (Barberá, 2025). For instance, user queries may inadvertently include sensitive information, and developers may implement retrieval systems lacking adequate safeguards. Organizational practices, such as insufficient data management or weak access controls, further exacerbate vulnerabilities (Holt et al., 2021). STS theory facilitates an understanding of privacy risks as emergent properties of complex interactions among AI models, retrieval systems, and human users, rather than isolated technical issues. In RAG systems, the interplay between social and technical factors is evident in multiple aspects. Retrieval processes depend on curated datasets that vary in data governance maturity, while user interactions are often unpredictable, complicating query resolution. These dynamics can result in unintended outcomes, such as the exposure of confidential information or the execution of harmful instructions (Bansal et al., n.d.). Accordingly, this study employs STS theory to define privacy risks as consequences arising from the alignment or misalignment among technological design, user intentions, and organizational safeguards.

Complementing this, Privacy by Design (PbD) provides a proactive and normative framework for embedding privacy safeguards into system development. Privacy by Design (PbD) requires incorporating technical and organizational measures in the earliest stages of system development and processing operations to protect privacy and uphold data protection principles actively. This method makes sure that privacy is an essential part of the design rather than something considered later (Ježová, 2020). In the context of RAG systems, this involves implementing privacy protections at multiple stages, including data ingestion, indexing, retrieval, and output generation.

2.7 Literature Gap

The current research on RAG systems primarily examines separate privacy problems, such as prompt injection and data leakage, without considering them across the entire system lifecycle. There is no unified framework that brings together attack analysis, privacy measurement, and risk reduction. Moreover, socio-technical views are underexplored. Most studies focus on technical controls and neglect user behavior and organizational governance. This restricts the creation of effective, real-world privacy solutions.

3 MATERIALS AND METHODS

3.1 Search Strategy

A systematic literature review was conducted to identify, evaluate, and synthesize existing research on privacy risks and mitigation strategies in Retrieval-Augmented Generation systems. The review adhered to established principles to ensure validity, transparency, and reproducibility (Eden et al., 2011). The search encompassed multiple academic and technical databases, including IEEE Xplore, SpringerLink, ScienceDirect, arXiv, and Google Scholar, to collect both peer-reviewed research and high-quality preprints relevant to this rapidly evolving field.

A structured Boolean search strategy was used to balance sensitivity and specificity (Bramer et al., 2018). The search combined terms related to RAG architectures with concepts tied to privacy and security. This method ensured that the research covered studies on both system-level vulnerabilities and privacy-preserving techniques. The study limited the search to publications from 2021 to 2026 to reflect the emergence and rapid development of RAG systems and associated privacy concerns.

Inclusion criteria were set to ensure relevance and quality. Only peer-reviewed articles and reputable preprints that discussed RAG systems or LLM-based retrieval architectures and examined privacy risks, attacks, or mitigation methods were used in the study. Studies that lacked technical depth, empirical analysis, or relevance to retrieval-based AI systems, were not in English, or were published before 2021 were excluded. Thus, the final collection made significant contributions to the understanding of privacy in RAG systems.

Table 1: Inclusion/exclusion criteria

Criterion	Included	Excluded
Source type	Peer-reviewed articles and reputable preprints on RAG systems or LLM-based retrieval architectures	Studies that were not peer-reviewed, not reputable preprints, or not focused on RAG/retrieval-based AI
Topical focus	Privacy risks, attacks, or mitigation methods related to RAG systems	Studies lacking relevance to privacy in retrieval-based AI systems
Technical quality	Papers with technical depth and empirical analysis	Studies with limited technical depth or no empirical analysis
Language	English-language studies	Non-English studies
Publication date	Published from 2021 onward	Published before 2021

3.2 Study Selection Using PRISMA Framework

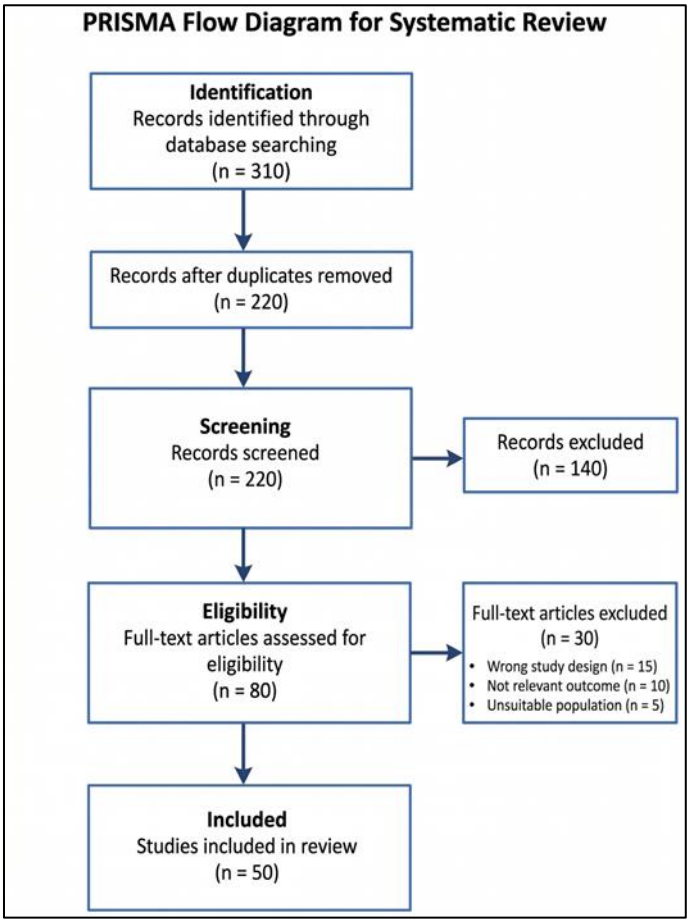

Figure 2: Prisma Flow Chart (Self-created)

The study selection process followed the PRISMA framework. The initial database search identified 310 records using a structured Boolean search expression:

("RAG" OR "retrieval augmented generation" OR "LLM retrieval systems") AND ("privacy" OR "security" OR "data protection" OR "information leakage")

After removing 90 duplicate entries, 220 unique records remained for screening. Then, titles and abstracts were screened for relevance, resulting in the exclusion of 140 studies that did not directly address RAG systems or privacy issues.

The full text of the remaining 80 studies was assessed to evaluate their quality and relevance to the research objectives. During this process, 30 papers were excluded for insufficient focus on privacy, insufficient depth, or failure to address retrieval-based AI architectures. The final dataset comprised 50 studies, which formed the basis for the systematic synthesis and thematic analysis in the results and discussion sections.

3.3 Data Analysis Technique

A two-stage synthesis approach was used to examine the selected studies. In the first stage, descriptive analysis mapped publication trends, research areas, types of RAG architectures studied, and the nature of privacy threats discussed. This overview showed how research in this area has evolved and highlighted key areas of focus in the literature.

In the second stage, thematic analysis was used to combine findings from the studies. The analysis focused on the main categories of privacy attacks, privacy risk measurement techniques, and mitigation strategies. By cross-analyzing, relationships between identified risks and their corresponding controls were explored to find patterns, overlaps, and gaps. This method enabled a thorough comparison of how different studies understand and address privacy challenges in RAG systems.

4 RESULTS

4.1 Introduction

This section presents the findings from a systematic review of 50 peer-reviewed sources addressing the research objectives. The analysis integrates Socio-Technical Systems Theory, which conceptualizes RAG systems as interdependent technical artifacts and social practices involving users, organizations, and regulatory contexts, with Privacy by Design (PbD) principles. These principles underscore the necessity of embedding privacy protections at every stage of system design. This perspective frames RAG as a socio-technical ecosystem in which privacy vulnerabilities arise from the interplay among data flows, human actors, and institutional governance. The discussion is organized thematically around the research objectives and concludes with the proposal of the Integrated Privacy-Preserving RAG Framework (IPRAG), which operationalizes the socio-technical and PbD foundations.

4.2 Theme 1: Identification and Categorization of Privacy Attacks in RAG Systems

The literature identifies five main categories of privacy attacks that threaten RAG systems. Membership Inference Attacks pose a significant vulnerability, allowing attackers to determine whether specific documents are in the retrieval database. Anderson et al. (2024) show that attackers can efficiently make these inferences using carefully designed prompts in both black-box and grey-box settings across various generative models. In addition, Li et al. (2024) introduce the Semantic Similarity-based Membership Inference Attack (S2MIA), which leverages the semantic relationship between query samples and RAG-generated outputs. This method outperforms five established baselines, even when faced with three representative defenses. Arzanipour et al. (2025) provide formal mathematical definitions of document-level membership inference and classify it as a key threat within a broader classification of RAG-specific risks.

Data Poisoning Attacks form the second category, in which malicious actors inject harmful content to alter system behavior. Wang et al. (2025) introduce PoisonedRAG, demonstrating that small amounts of poisoned text can yield targeted outputs, and provide the first complete security analysis of RAG pipelines. Gummadi et al. (2024) classify data poisoning, model manipulation, and privacy leakage as critical security issues, while Zhou et al. (2025) emphasize the need for formal security guarantees rather than makeshift defenses.

Adversarial and Injection Attacks include techniques like prompt injection, jailbreaking, and multilingual manipulation. ElSaify and Baderelden (2025) present a detailed taxonomy in this area, noting that Arabic-English code-switching effectively bypasses safety filters. Ward and Harguess (2025) connect adversarial query manipulation to risk-management frameworks, while Liang et al. (2025) contextualize these threats through the SafeRAG benchmark. They distinguish among silver noise, inter-context conflict, soft advertisements, and white denial-of-service attacks.

Privacy Leakage Attacks and Structured Data Extraction complete the classification. Zeng et al. (2024a) highlight a paradox in which RAG systems unintentionally reveal retrieval-database contents while reducing leaks of the underlying LLM training data. Zhang et al. (2025) extend these observations to multimodal contexts. Guan et al. (2025)

highlight inconsistencies in the safeguarding of protected health information in clinical settings. Liu et al. (2025) show that Graph RAG architectures may reduce raw-text leakage but increase the risk of structured-entity and relationship extraction. These categories illustrate how socio-technical interactions, including retrieval methods, generation processes, and human oversight, can elevate privacy risks unless PbD principles are incorporated from the beginning.

4.3 Theme 2: Analysis of Privacy Risks Across RAG Pipeline Stages

Privacy risks appear in different ways across the storage, retrieval, and generation stages of the RAG pipeline, supporting the socio-technical interdependence suggested by the theoretical framework. At the storage stage, Zhou et al. (2025a) emphasize the growing vulnerability of proprietary knowledge bases to breaches as RAG use increases. Ammann et al. (2025) map attack surfaces from preprocessing to infrastructure, while Yu and Sato (2024) suggest decentralized storage to avoid single points of failure.

Risks during the retrieval stage focus on query exposure and embedding leakage. Wang et al. (2025) introduce PIR-RAG, which utilizes lattice-based Private Information Retrieval with semantic clustering to protect user queries from service providers. He et al. (2025) propose shifting embedding spaces to reroute privacy-leaking queries. Bodea et al. (2026) categorize these risks within a thorough privacy taxonomy.

Generation-stage threats arise when the LLM synthesizes retrieved content. Arzanipour et al. (2025) formalize the ways in which this integration creates new sources of vulnerability. Koga et al. (2024) and Grislain (2024) address these issues through differentially private token generation, demonstrating that this approach can operate within practical privacy budgets.

Cross-stage analyses reinforce the socio-technical perspective. Guan et al. (2025) provide a comprehensive, pipeline-based framework encompassing storage, transmission, retrieval, and generation, with a particular focus on healthcare applications. Oche et al. (2025) and Sharma (2025) highlight ongoing trade-offs between retrieval accuracy, generation flexibility, and privacy, emphasizing the need for PbD integration throughout the entire socio-technical ecosystem.

4.4 Theme 3: Evaluation of Existing Privacy Measurement and Evaluation Techniques

Robust evaluation is crucial for implementing PbD within socio-technical RAG systems. Formal frameworks include Liang et al.'s (2025) SafeRAG benchmark, which examines 14 components across four attack categories, and Zhou et al.'s (2024) six-dimensional trustworthiness evaluation, which covers factuality, robustness, fairness, transparency, accountability, and privacy. Yu et al. (2024) emphasize the evaluation challenges posed by RAG's mixed architecture.

In healthcare, Binwal and Chopra (2024) estimate a 3% risk of sensitive data exposure and a 2.7% risk of unintentional disclosure of personal information in unprotected systems. Neha et al. (2025) recommend domain-specific metrics, such as FactScore, RadGraph-F1, and MED-F1, while Amugongo et al. (2025) analyze 30 studies to evaluate diagnostic, summarization, and question-answering tasks using PRISMA guidelines.

Performance-privacy trade-offs have been quantified by Zeng et al. (2024), Koga et al. (2024), and Binwal and Chopra (2024), who report that differential privacy techniques can reduce exposure by up to 97% while maintaining 92% of baseline performance. Benchmark development by Barnett et al. (2024) and Sharma (2025) emphasizes that RAG robustness develops operationally rather than through static design. This reinforces the need for ongoing, context-aware evaluation.

4.5 Theme 4: Comparison of Mitigation Strategies (Technical and Governance Controls)

Mitigation strategies include technical, algorithmic, and governance controls, each assessed for effectiveness, overhead, and alignment with PbD. Encryption-based controls offer verifiable security. Zhou et al. (2025) propose the SAG framework, which employs full encryption of both content and embeddings before storage. Zhou et al. (2025a) extend this to privacy-aware isolated retrieval with formal proofs.

Privacy-preserving algorithmic methods include Zeng et al.'s (2024) SAGE synthetic-data approach and He et al.'s (2025) PRESS embedding-space shifting. Mao et al. (2025) combine federated learning with homomorphic encryption in FedE4RAG. Differential privacy methods (Koga et al., 2024; Grislain, 2024) allocate privacy budgets selectively to sensitive tokens.

Governance controls supplement technical measures. Panda and Mukherjee (2025) support adopting zero-trust architectures to ensure compliance with the GDPR and the CCPA. Binwal and Chopra (2024) report a 40% reduction in biased outputs through the use of ethical subroutines, while Guan et al. (2025) emphasize that current safeguards are inadequate in clinical settings.

Hybrid approaches show better results. Al Masoud et al. (2026) report a 58% improvement in privacy scores with SD-RAG, while Ammann et al. (2025) combine RAG-specific security with industry standards. Quantitative assessments by Anderson et al. (2024) and Ward and Harguess (2025) confirm that template-based instructions, adversarial training, and real-time monitoring are effective when placed within a socio-technical governance framework.

4.6 Theme 5: Proposal of an Integrated Framework for Privacy-Preserving RAG Systems

The review leads to the Integrated Privacy-Preserving RAG Framework (IPRAG), which operationalizes Socio-Technical Systems Theory and PbD by embedding privacy across five architectural layers and following a three-phase implementation protocol.

- Layer 1 (Threat Intelligence & Risk Assessment) incorporates formal taxonomies (Arzanipour et al., 2025; Bodea et al., 2026) that cover membership inference, data poisoning, leakage, adversarial manipulation, and structured extraction.
- Layer 2 (Pipeline-Stage Protection) uses storage-level encryption (Zhou et al., 2025; Zhou et al., 2025a), retrieval-level private information retrieval (Wang et al., 2025), and generation-level differential privacy (Koga et al., 2024).
- Layer 3 (Multi-Modal Privacy Measurement) integrates SafeRAG benchmarking (Liang et al., 2025), healthcare metrics (Binwal & Chopra, 2024), and six-dimensional trustworthiness evaluation (Zhou et al., 2024).
- Layer 4 (Adaptive Defense Orchestration) coordinates technical controls, achieving a 97% reduction in exposure through SAGE and differential privacy, along with governance controls that reduce bias by 40% and ensure regulatory compliance.
- Layer 5 (Domain-Specific Customization) adapts the framework for healthcare (Guan et al., 2025; Amugongo et al., 2025), education (Li et al., 2025), and business contexts.

The three-phase protocol: Assessment & Planning, Defense Architecture Design, and Deployment & Monitoring, aims for measurable results: less than 1% successful attacks, over 95% data protection, more than 90% performance retention, and 100% regulatory compliance. Future pathways include agentic RAG (Singh et al., 2025), adaptive architectures (Sharma, 2025; Oche et al., 2025), and multimodal extensions to keep the framework responsive to changing socio-technical conditions.

In summary, the thematic analysis demonstrates a progression from identifying discrete vulnerabilities to developing a comprehensive framework. The Integrated Privacy-Preserving RAG Framework (IPRAG) synthesizes these advancements into a unified, Privacy by Design-compliant, and socio-technically informed solution that balances privacy, utility, and regulatory compliance. This approach advances both academic understanding and the practical implementation of secure RAG systems.

5 DISCUSSION

The preceding analysis demonstrates significant advancements in RAG privacy and security research, progressing from the identification of isolated attacks (Theme 1) to the mapping of pipeline-stage risks (Theme 2), the application of rigorous evaluation methods (Theme 3), the comparison of mitigation strategies (Theme 4), and the development of comprehensive frameworks (Theme 5). This progression aligns with the five research objectives. It is grounded in Socio-Technical Systems Theory, which conceptualizes RAG as an interconnected system of technical components, human actors, organizational practices, and regulatory environments. The analysis also draws upon Privacy by Design (PbD), which advocates for proactive privacy measures rather than reactive solutions.

The proposed Integrated Privacy-Preserving RAG Framework (IPRAG) represents the culmination of these 50 studies. As shown in Figure 2, IPRAG comprises five layers and is supported by a three-phase implementation process. This ensures that privacy is not just an extra feature, but a fundamental quality of the socio-technical system. The framework sets measurable goals of fewer than 1% successful attacks, over 95% data protection, over 90% performance retention, and 100% regulatory compliance. It also allows for adjustments based on specific domains and future developments, such as agentic and multi-modal RAG.

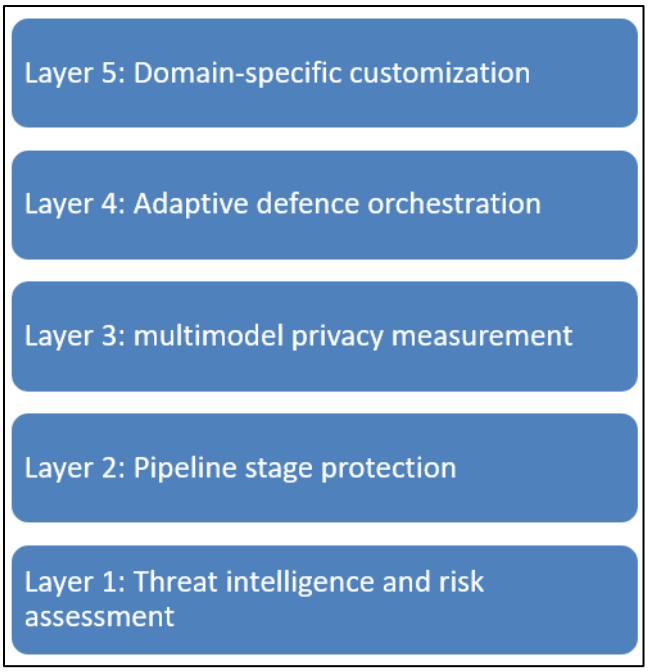

Figure 3: IPRAG Framework

Figure 2 illustrates the IPRAG framework. This model translates the theoretical concepts into practice by placing Threat Intelligence & Risk Assessment and Pipeline-Stage Protection at the base. It adds Adaptive Defense Orchestration and Domain-Specific Customization to enable flexible, context-aware responses that align with Socio-Technical Systems Theory. PbD is achieved through the clear integration of privacy measurement and defense coordination at all architectural levels and in every phase of implementation.

6 CONCLUSION

6.1 Summary of the Findings

This study synthesized 50 sources to explore security and privacy challenges in Retrieval-Augmented Generation (RAG) systems. The analysis identified five main categories of privacy attacks: membership inference attacks, data poisoning attacks, adversarial and injection attacks, privacy leakage attacks, and structured data extraction. These attacks exploit weaknesses in the RAG pipeline. Notably, RAG systems may leak content from the retrieval database while still protecting the underlying LLM training data (Zeng et al., 2024a).

The risks were found to be specific to different stages but still interconnected. They impact the storage phase (breaches of proprietary knowledge bases), the retrieval phase (exposure of queries), and the generation phase (vulnerabilities in synthesis). Current privacy measurement techniques include formal benchmarks like SafeRAG (Liang et al., 2025) and six-dimensional trust evaluations (Zhou et al., 2024). Healthcare-specific metrics indicate a 3% risk of sensitive data exposure in unprotected systems (Binwal & Chopra, 2024). The study showed that technical controls (such as encryption, synthetic data, and differential privacy), governance controls (such as zero-trust and regulatory compliance), and hybrid strategies result in measurable improvements, including a 97% reduction in exposure and a 58% increase in privacy scores.

The synthesis led to the development of the Integrated Privacy-Preserving RAG Framework (IPRAG), as shown in Figure 2. This framework combines these elements into a practical architecture.

Linking with the Objectives

The findings directly address the study's five research objectives.

- Objective 1, which focused on identifying and categorizing privacy attacks, was met by developing a structured taxonomy grounded in formal threat models (Arzanipour et al., 2025; Bodea et al., 2026).
- Objective 2, which aimed to analyze privacy risks across different pipeline stages, was fulfilled by mapping vulnerabilities and interdependencies from a socio-technical perspective (Guan et al., 2025; Zhou et al., 2025a).

- Objective 3, which involved evaluating privacy measurement techniques, was achieved through a critical review of benchmarks, domain-specific metrics, and the trade-offs between performance and privacy (Liang et al., 2025; Binwal & Chopra, 2024).
- Objective 4 compared mitigation strategies, evaluating technical, governance, and hybrid controls with quantitative evidence of their effectiveness (Zhou et al., 2025; Al Masoud et al., 2026).
- Objective 5 proposed an integrated framework, realized through the IPRAG framework. This framework applies Socio-Technical Systems Theory and Privacy by Design principles across five layers and a three-phase implementation protocol. Overall, the study shows a clear progression from identifying vulnerabilities to developing comprehensive, well-grounded solutions.

Recommendations

- Practical Recommendations

Organizations using RAG systems should adopt the IPRAG framework as standard practice. They should prioritize full pre-storage encryption (Zhou et al., 2025; 2025a), use differential privacy during generation (Koga et al., 2024), and customize approaches for specific domains, especially healthcare (Guan et al., 2025). Deployment should follow the three-phase protocol to meet target metrics: less than 1% successful attacks, over 95% data protection, over 90% performance retention, and full regulatory compliance.

- Research Recommendations

Future empirical studies should validate IPRAG through real-world case studies and long-term evaluations. Researchers should focus on agentic RAG (Singh et al., 2025), multimodal privacy risks (Zhang et al., 2025), and Graph RAG-structured extraction vulnerabilities (Liu et al., 2025). The development of open-source IPRAG toolkits is encouraged to improve reproducibility and community feedback.

REFERENCES

[1] Al Masoud, A., Arazzi, M. & Nocera, A. (2026). 'SD-RAG: A Prompt-Injection-Resilient Framework for Selective Disclosure in Retrieval-Augmented Generation ', *arXiv preprint arXiv:2601.11199*. doi:10.48550/arXiv.2601.11199.

[2] Amazon (2024).*What is RAG? - Retrieval-Augmented Generation Explained – AWS*—[online] Amazon Web Services, Inc. Available at: https://aws.amazon.com/what-is/retrieval-augmented-generation/.

[3] Ammann, L., Ott, S., Landolt, C.R. & Lehmann, M. (2025). 'Securing RAG: A Risk Assessment and Mitigation Framework', in Swiss *Conference on Data Science*. https://doi.org/10.48550/arXiv.2505.08728.

[4] Amugongo, L.M. et al. (2025). 'Retrieval augmented generation for large language models in healthcare: A systematic review ', *PLOS Digital Health*, 4(6), p. e0000877. doi: 10.1371/journal.pdig.0000877.

[5] Amugongo, L.M., Mascheroni, P., Brooks, S., Doering, S., & Seidel, J. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review.*PLOS Digital Health*, 4(6), p.e0000877. doi: https://doi.org/10.1371/journal.pdig.0000877.

[6] Anderson, M., Amit, G., & Goldsteen, A. (2024). 'Is My Data in Your Retrieval Database? Membership Inference Attacks Against Retrieval Augmented Generation', in International *Conference on Information Systems Security and Privacy*(arXiv preprint arXiv:2405.20446). doi:10.48550/arXiv.2405.20446.

[7] Arzanipour, A. et al. (2025). 'RAG Security and Privacy: Formalizing the Threat Model and Attack Surface', arXiv.org. https://doi.org/10.48550/arXiv.2509.20324.

[8] Bansal, G., Wortman Vaughan, J., Amershi, S., Horvitz, E., Fourney, A., Mozannar, H., Dibia, V., & Weld, D. (n.d.).*Challenges in Human-Agent Communication*. [online] Available at: https://www.microsoft.com/en-us/research/wp-content/uploads/2024/12/HCAI_Agents.pdf.

[9] Barberá, I. (2025).*Large Language Models (LLMs) SUPPORT POOL OF EXPERTS PROGRAMME*. [online] Available at: https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf.

[10] Barnett, S. et al. (2024). 'Seven Failure Points When Engineering a Retrieval Augmented Generation System', in 2024 *IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN)*(arXiv preprint arXiv:2401.05856). doi:10.48550/arXiv.2401.05856.

[11] Bazzi, S.W. & Gaith, M. (2024).*The Wonders of RAG: Streamlining Knowledge with Advanced Techniques, Systematic Literature Review Report*. [online] doi: https://doi.org/10.13140/RG.2.2.26238.40001.

[12] Binwal, A. & Chopra, P. (2024). 'Privacy and Regulatory Compliance in Retrieval-Augmented Generation Models for AGI Systems', *International Journal for Multidisciplinary Research,* 6(6), https://doi.org/10.36948/ijfmr.2024.v06i06.30421.

[13] Binwal, A. & Chopra, P. (2024). Privacy and Regulatory Compliance in Retrieval-Augmented Generation Models for AGI Systems. *International Journal for Multidisciplinary Research*, 6(6), November-December. Available at: https://www.ijfmr.com/research-paper.php?id=30421 (Accessed: 1 April 2026).

[14] Bodea, A. et al. (2026). 'SoK: Privacy Risks and Mitigations in Retrieval-Augmented Generation Systems',*arXiv preprint arXiv:2601.03979*. doi:10.48550/arXiv.2601.03979.

[15] Bramer, W.M., De Jonge, G.B., Rethlefsen, M.L., Mast, F., & Kleijnen, J. (2018). A Systematic Approach to Searching: An Efficient and Complete Method to Develop Literature Searches.*Journal of the Medical Library Association*, 106(4), pp.531–541. doi: https://doi.org/10.5195/jmla.2018.283.

[16] Cheng, M. et al. (2025). 'A Survey on Knowledge-Oriented Retrieval-Augmented Generation', arXiv *preprint arXiv:2503.10677*. doi:10.48550/arXiv.2503.10677.

[17] Cuconasu, F. et al. (2024). 'The Power of Noise: Redefining Retrieval for RAG Systems', in Annual *International ACM SIGIR Conference on Research and Development in Information Retrieval*(arXiv preprint arXiv:2401.14887). doi:10.48550/arXiv.2401.14887.

[18] Eden, J., Levit, L., Berg, A., & Morton, S. (2011).*Standards for Initiating a Systematic Review*. [online]*www.ncbi.nlm.nih.gov*. National Academies Press (US). Available at: https://www.ncbi.nlm.nih.gov/books/NBK209515/.

[19] ElSaify, B. & Baderelden, M. (2025). 'Adversarial and Multilingual Threats in Retrieval-Augmented Generation: From Prompt Injection to Model Exploitation', in 2025 *2nd International Generative AI and Computational Language Modeling Conference (GACLM)*. https://doi.org/10.1109/GACLM67198.2025.11231998.

[20] Gao, Y. et al. (2023). 'Retrieval-Augmented Generation for Large Language Models: A Survey ', *arXiv preprint arXiv:2312.10997*. doi:10.48550/arXiv.2312.10997.

[21] Gorejena, K., Mavetera, N. & Velempini, M. (2016). A critique and potency of socio-technical systems theory: a quest for broadband growth and penetration.*Public and Municipal Finance*, 5(2), pp. 7–19. doi:10.21511/pmf.5(2).2016.01

[22] Grislain, N. (2024). 'RAG with Differential Privacy', in Conference *on Algebraic Informatics*. https://doi.org/10.48550/arXiv.2412.19291

[23] Gu, J. (2025). 'A Research of Challenges and Solutions in Retrieval Augmented Generation (RAG) Systems', *Highlights in Science, Engineering and Technology*. https://doi.org/10.54097/364hex16

[24] Guan, S. et al. (2025). 'Privacy Challenges and Solutions in Retrieval-Augmented Generation-Enhanced LLMs for Healthcare Chatbots: A Review of Applications, Risks, and Future Directions', arXiv.org. https://doi.org/10.48550/arXiv.2511.11347

[25] Gummadi, V. et al. (2024). 'Enhancing Communication and Data Transmission Security in RAG Using Large Language Models', in International *Conference on Signals and Electronic Systems*. https://doi.org/10.1109/ICSES63445.2024.10763024

[26] Guo, Z. et al. (2024). 'LightRAG: Simple and Fast Retrieval-Augmented Generation', in Conference *on Empirical Methods in Natural Language Processing*(arXiv preprint arXiv:2410.05779). doi:10.48550/arXiv.2410.05779. https://doi.org/10.48550/arXiv.2511.11347

[27] Gupta, S., Ranjan, R. & Singh, S.N. (2024). 'A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions', *arXiv preprint arXiv:2410.12837*. doi:10.48550/arXiv.2410.12837.

[28] Han, H. et al. (2024). 'Retrieval-Augmented Generation with Graphs (GraphRAG)', arXiv.org. https://doi.org/10.48550/arXiv.2501.00309

[29] He, J. et al. (2025) 'PRESS: Defending Privacy in Retrieval-Augmented Generation via Embedding Space Shifting', in IEEE *International Conference on Acoustics, Speech, and Signal Processing*. https://doi.org/10.1109/ICASSP49660.2025.10887843

[30] He, L., Tang, P., Zhang, Y., Zhou, P., and Su, S. (2025). Mitigating privacy risks in Retrieval-Augmented Generation via locally private entity perturbation.*Information Processing & Management*, [online] 62(4), p.104150. doi: https://doi.org/10.1016/j.ipm.2025.104150.

[31] Holt, J., Baker, J., Blake, J., Barnes, J., & Israel, O. (2021).*Security Vulnerabilities in Database-as-a-Service Platforms and Their Impact on Data Confidentiality*. [online] Available at: https://www.researchgate.net/publication/398737741_Security_Vulnerabilities_in_Database_as_a_Service_Platforms_and_Their_Impact_on_Data_Confidentiality.

[32] James, A., Trovati, M., & Bolton, S. (2025). 'Retrieval-Augmented Generation to Generate Knowledge Assets and Creation of Action Drivers', *Applied Sciences*. https://doi.org/10.3390/app15116247

[33] Ježová, D. (2020). Principle of Privacy by Design and Privacy by Default. pp.127–139. doi: https://doi.org/10.18485/iup_rlr.2020.ch10.

[34] Klesel, M. & Wittmann, H.F. (2025). Retrieval-Augmented Generation (RAG).*Business & Information Systems Engineering*. doi: https://doi.org/10.1007/s12599-025-00945-3.

[35] Koga, T., Wu, R. & Chaudhuri, K. (2024). 'Privacy-Preserving Retrieval Augmented Generation with Differential Privacy', arXiv.org. https://doi.org/10.48550/arXiv.2412.04697

[36] Li, Y. et al. (2024). 'Generating Is Believing: Membership Inference Attacks against Retrieval-Augmented Generation', in IEEE *International Conference on Acoustics, Speech, and Signal Processing*. https://doi.org/10.48550/arXiv.2406.19234

[37] Li, Z. et al. (2025). 'Retrieval-Augmented Generation for Educational Application: A Systematic Survey', *Computers and Education: Artificial Intelligence*. https://doi.org/10.1016/j.caeai.2025.100417.

[38] Liang, X. et al. (2025). 'SafeRAG: Benchmarking Security in Retrieval-Augmented Generation of Large Language Model', in Annual *Meeting of the Association for Computational Linguistics*. https://doi.org/10.48550/arXiv.2501.18636

[39] Liu, J., Zhang, J. & Wang, S. (2025). 'Exposing Privacy Risks in Graph Retrieval-Augmented Generation', arXiv.org. https://arxiv.org/abs/2508.17222

[40] Malali, N. (2025). The Role of Retrieval-Augmented Generation (RAG) in Financial Document Processing: Automating Compliance and Reporting.*International Journal of Management Technology*, 12(3), pp.26–46. doi: https://doi.org/10.37745/ijmt.2013/vol12n32646.

[41] Mao, Q. et al. (2025). 'Privacy-Preserving Federated Embedding Learning for Localized Retrieval-Augmented Generation', arXiv.org. https://arxiv.org/abs/2504.19101

[42] Neha, F., Bhati, D. & Shukla, D.K. (2025). 'Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review', *Applied Informatics*. https://doi.org/10.3390/ai6090226

[43] Ni, B. et al. (2025). 'Towards Trustworthy Retrieval Augmented Generation for Large Language Models: A Survey', arXiv.org. https://arxiv.org/abs/2502.06872

[44] Oche, A.J. et al. (2025). 'A Systematic Review of Key Retrieval-Augmented Generation (RAG) Systems: Progress, Gaps, and Future Directions', arXiv.org. https://arxiv.org/abs/2507.18910

[45] Ozman, F.M. (2025). Systematic literature review on the evolution of human-machine collaboration in the workplace.*World Journal of Advanced Engineering Technology and Sciences*, 15(3), pp.1981–1999. doi: https://doi.org/10.30574/wjaets.2025.15.3.1043.

[46] Panda, M. & Mukherjee, S. (2025). 'Enhancing Privacy and Security in RAG-Based Generative AI Applications', *AI, Machine Learning and Applications advances 2025*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5162147

[47] Panda, M. & Mukherjee, S. (2025). Advancing Privacy and Security in Generative AI-Driven Rag Architectures: A Next-Generation Framework.*International Journal of Artificial Intelligence & Applications*, 16(2), pp.15–24. doi: https://doi.org/10.5121/ijaia.2025.16203.

[48] Peng, B. et al. (2024). 'Graph Retrieval-Augmented Generation: A Survey', *ACM Transactions on Information Systems*. https://arxiv.org/abs/2408.08921.

[49] Polan, S. (2025). 'Retrieval-Augmented Generation: Architecture, Techniques, and Evaluations', *Journal of Modern Technology and Engineering*, 10(1), pp. 42–56. doi: 10.62476/jmte.10142

[50] Sawarkar, K., Mangal, A. & Solanki, S. (2024). 'Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers', in Conference *on Multimedia Information Processing and Retrieval*. https://arxiv.org/abs/2404.07220

[51] Shahade, A.K. & Deshmukh, P.V. (2024). 'Enhancing Natural Language Processing: A Comprehensive Review of Retrieval Augmented Generation', in International *Conference on Signals and Electronic Systems*. https://doi.org/10.1109/ICSES63445.2024.10763224.

[52] Sharma, C. (2025). 'Retrieval-Augmented Generation: A Comprehensive Survey of Architectures, Enhancements, and Robustness Frontiers', arXiv.org. https://arxiv.org/html/2506.00054v1.

[53] Singh, A. et al. (2025). 'Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG', arXiv.org. https://arxiv.org/abs/2501.09136.

[54] Soman, S. & Roychowdhury, S. (2024). 'Observations on Building RAG Systems for Technical Documents', in Tiny *Papers @ ICLR*. https://arxiv.org/abs/2404.00657

[55] Wang, B. et al. (2025). 'PIR-RAG: A System for Private Information Retrieval in Retrieval-Augmented Generation', arXiv.org. https://arxiv.org/abs/2509.21325.

[56] Wang, C. et al. (2025). 'Retrieval-Augmented Generation: A Survey of Security Challenges and Countermeasures', in 2025 *11th IEEE International Conference on Privacy Computing and Data Security (PCDS)*. https://doi.org/10.1109/PCDS65695.2025.00037

[57] Ward, C.M. & Harguess, J.D. (2025). 'Adversarial threat vectors and risk mitigation for retrieval-augmented generation systems', in Defense *+ Security*. https://arxiv.org/abs/2506.00281.

[58] Wu, S. et al. (2024). 'Retrieval-Augmented Generation for Natural Language Processing: A Survey', arXiv.org. https://arxiv.org/abs/2407.13193

[59] Yu, H. et al. (2024). 'Evaluation of Retrieval-Augmented Generation: A Survey', arXiv.org. https://arxiv.org/abs/2405.07437

[60] Yu, J. & Sato, H. (2024). 'DeRAG: Decentralized Multi-Source RAG System with Optimized Pyth Network', in International *Symposium on Image and Signal Processing and Analysis*. https://www.ieee-ai-for-science.org/2024/ispa/ispa-2024-accepted.html.

[61] Zeng, S. et al. (2024). 'Mitigating the Privacy Issues in Retrieval-Augmented Generation (RAG) via Pure Synthetic Data', in Conference *on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/2406.14773

[62] Zeng, S. et al. (2024a). 'The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG)', in Annual *Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2024.findings-acl.267.

[63] Zhang, J. et al. (2025). 'Beyond Text: Unveiling Privacy Vulnerabilities in Multi-modal Retrieval-Augmented Generation', in Conference *on Empirical Methods in Natural Language Processing*. https://doi.org/10.48550/arXiv.2505.13957

[64] Zhao, P. et al. (2024). 'Retrieval-Augmented Generation for AI-Generated Content: A Survey', *Data Science and Engineering*(arXiv preprint arXiv:2402.19473). doi:10.48550/arXiv.2402.19473.

[65] Zhou, P., Feng, Y., & Yang, Z. (2025). 'Provably Secure Retrieval-Augmented Generation', arXiv.org. https://arxiv.org/abs/2508.01084.

[66] Zhou, P., Feng, Y., & Yang, Z. (2025a). 'Privacy-Aware RAG: Secure and Isolated Knowledge Retrieval', arXiv.org.

[67] Zhou, Y. et al. (2024). 'Trustworthiness in Retrieval-Augmented Generation Systems: A Survey', arXiv *preprint arXiv:2409.10102*. doi:10.48550/arXiv.2409.10102.