

# Leveraging RAG for Privacy Violation Detection and Explainability

Stefano Locci  
*Department of Computer Science*  
*University of Turin*  
Turin, Italy  
0009-0006-9725-2045

Davide Audrito  
*Law Department*  
*Autonomous University of Barcelona*  
Bellaterra, Spain  
0000-0002-9239-5358

Giovanni Livraga  
*Department of Computer Science*  
*University of Milan*  
Milan, Italy  
0000-0003-2661-8573

Marco Viviani  
*Department of Informatics, Systems, and Communication*  
*University of Milano-Bicocca*  
Milan, Italy  
0000-0002-2274-9050

Luigi Di Caro  
*Department of Computer Science*  
*University of Turin*  
Turin, Italy  
0000-0002-7570-637X

**Abstract**—In today’s digital landscape, users frequently share vast amounts of information, including confidential data, often without full awareness of the associated privacy risks. This scenario highlights the need for automated methods to identify sensitive information and alert users to such risks. Existing algorithmic solutions for detecting sensitive content typically require either human intervention (rule-based approaches) or labeled data (supervised learning), both of which can be costly and limiting. In this paper, we propose a framework based on Retrieval-Augmented Generation (RAG) to classify privacy-sensitive content while providing contextual explanations. We employed the state-of-the-art generative Large Language Model (LLM) GPT-4o, with Information Retrieval models BM25 and FAISS, enhancing both detection accuracy and explainability. Our method utilizes a curated Knowledge Base of scientific literature on privacy and confidentiality to retrieve contextually relevant information, which is then used to guide the classification process and generate explanations. Experimental evaluations on a real-world dataset (Enron Email Dataset) demonstrate that RAG-based approaches significantly outperform the zero-shot baseline, with BM25 showing the highest performance. This tool is designed to serve end-users, by mitigating risks before data sharing, by enabling proactive monitoring of privacy violations.

**Index Terms**—Privacy, Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), Information Retrieval (IR), Knowledge Bases (KBs)

## I. INTRODUCTION

The vast and rapid exchange of information in the current online landscape presents both opportunities and challenges, particularly concerning data privacy. Users often unknowingly share confidential information, making it essential to implement effective mechanisms that can identify such data and alert users to potential privacy violations. Current algorithmic approaches for detecting confidential information often rely on rule-based methods or supervised learning techniques, which can be resource-intensive and may require extensive human intervention or labeled datasets. As a result, there is a pressing need for more efficient, automated solutions that can address these challenges in real time.

In this context, we introduce a novel framework that leverages *Retrieval-Augmented Generation* (RAG) to advance the detection of privacy-sensitive information through the use of *Large Language Models* (LLMs). Our approach draws on a specialized *Knowledge Base* (KB) consisting of scientific literature focused on privacy and confidentiality. Two well-established *Information Retrieval* (IR) models are employed: BM25 (for sparse retrieval) and FAISS (for dense retrieval). First, the proposed solution addresses the task of classifying text for privacy violations by identifying potentially sensitive information within unstructured text. Furthermore, it enhances the generative model’s ability to explain privacy violations through natural language, making insights more accessible and understandable for users. This dual approach enhances both the detection accuracy and the interpretability of privacy-related risks. Our approach is evaluated by considering the performance of the OpenAI GPT-4o model in both a zero-shot approach and in conjunction with the proposed RAG using the considered KB. Our results highlight the potential of the RAG system to improve online privacy in practical applications. The code is made publicly available on Github.<sup>1</sup>

In the following, we first review the related literature on automated privacy detection, explainability and RAG, discussing the limitations of existing approaches in addressing privacy concerns within unstructured data (Section II). We then present our methodology in detail, describing the pipeline, the retrieval models employed, and the dataset and KB creation process (Section III). Next, we outline the experimental setup used to evaluate our system, including the dataset, evaluation metrics, and specific configurations tested (Section IV). Finally, we discuss the results obtained, analyze the performance of our proposed framework, and highlight potential future directions to further enhance the detection and explainability of privacy-sensitive content (Sections V and VI).

<sup>1</sup><https://github.com/stefanolocci/RAG-for-Privacy-Violation-Detection>

## II. RELATED WORK

In this section, we review relevant literature on the main research areas involved in this work, namely confidentiality protection, explainability of automated systems, and Retrieval-Augmented Generation.

### A. Confidentiality protection

Given the vast amount of information processed daily by automated systems, privacy-preserving techniques are essential to safeguard user information confidentiality and protect against unauthorized access and misuse [1]. Despite the publication of multiple frameworks and regulations in recent years, such as GDPR, CCPA, and HIPAA to enforce data protection standards [2], maintaining data privacy in practice remains challenging, particularly when handling unstructured data. In this scenario, which encompasses emails, chat logs, multimedia content, etc., this is particularly complex due to the lack of predefined structure, making it more difficult to identify and protect sensitive information.

Among the various solutions proposed for protecting the confidentiality of unstructured data, various methodologies have been investigated. Rule-based methods utilize predefined patterns and vocabularies to identify sensitive content [3], [4]. While effective in certain (limited) scenarios, they lack the flexibility needed for broader applications. A more automated solution is represented by *Named Entity Recognition* (NER)-based methods, which use *Machine Learning* (ML) techniques to identify and categorize potential sensitive entities in text [5], [6]. However, these methods are usually limited to predefined categories and often restricted to a single domain. Approaches employing ML rely on datasets labeled according to the specific problem at hand [4], [7], necessitating a labeling process that encompasses all the challenges associated with that task. A promising approach is the use of *Recursive Neural Networks* (RNNs) for detecting sensitive information, as demonstrated by the TABOO engine [8]. This method utilizes RNNs to assign sensitivity scores by learning the semantic and syntactic structure of text documents, allowing it to capture complex contexts where sensitive content might not be explicitly identifiable through keywords or standard classification techniques. The RNN-based approach requires domain-specific labeled data and necessitates retraining when applied to new domains. Additionally, RNNs struggle with semantic ambiguity and subtle contextual cues, which can lead to misclassifications. To address these limitations, BERT-based approaches have been explored due to their ability to process unstructured text and capture contextual information more effectively [9]. Despite their effectiveness, BERT models often encounter challenges related to domain dependency. Since they are pretrained on general-purpose corpora, their performance in specialized domains may be suboptimal unless fine-tuned on large, domain-specific datasets. To overcome these challenges, recent frameworks have begun utilizing LLMs for tasks such as classification [10]. Notably, recent solutions based on RAG show promise in providing accurate domain knowledge and improving the clarity of explanations.

### B. Explainability

Explainability is crucial for systems that aim to protect privacy. When these systems flag something as potentially confidential, users and organizations need to understand the reasons behind this decision in order to coherently mitigate that risk. Traditional methods often simply label information as either private or not, without providing any explanation. However, recent advancements in the field of *Explainable Artificial Intelligence* (XAI) have highlighted the importance of providing clear and understandable explanations to build trust and transparency in automated systems [11]. Despite these advancements, many current XAI methods are limited to structured data or predefined rules. This restricts their applicability to unstructured text like emails or social media posts, and to very large datasets [12]. Unstructured data poses a well-known challenge for explainability due to its absence of predefined structure and the context-dependent nature of its content [13]. In scenarios involving private and sensitive information, this challenge becomes even more significant, as the sensitivity of data often hinges on the surrounding context.

### C. Retrieval-Augmented Generation Systems

In RAG-based systems, a retrieval process operates alongside a generative process: the retrieval model identifies relevant documents from an external KB, while the generative model leverages this information to produce more accurate outputs. RAGs have demonstrated high efficacy in tasks such as *Question Answering* (QA) [14] and misinformation detection [15], [16], utilizing sparse retrieval (e.g., BM25), dense retrieval (e.g., FAISS) [17], and hybrid approaches [15].

While RAG is primarily used to guide LLMs towards generating more contextualized and referenced content, thereby reducing hallucinations, it is employed in [18] to detect deceptive content. Building on this approach, we propose a novel RAG-based pipeline that assists LLMs in evaluating the risk of disclosing confidential information from a collection of documents, as detailed below.

## III. METHODOLOGY

Figure 1 illustrates the RAG framework considered in our work, further discussed below. In the initial phase, *each entry* of a given textual dataset (in our case, the Enron Mail Dataset detailed in Section III-B, specifically the body of the email) that requires classification for potential privacy violations is used as a *query* to search a KB containing *scientific evidence* (e.g., scientific articles, technical reports, etc.) on privacy-related topics. The KB is divided into *passages*, with each passage consisting of a *sentence*, and the top- $k$  passages relevant to the query are retrieved using an IR model leveraging BM25 for sparse retrieval and FAISS for dense retrieval. These  $k$  passages are then incorporated into an appropriate *prompt*, which guides the LLM in classifying the considered textual content (i.e., the email in our case) for privacy violations and explaining why it may pose a risk.

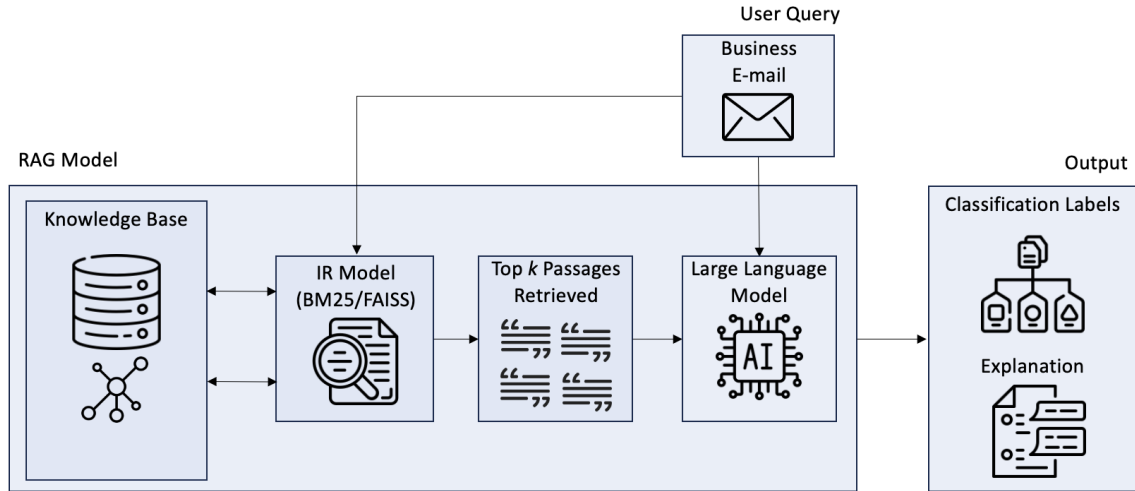


Fig. 1. The proposed pipeline for privacy violation detection using a **Retrieval-Augmented Generation (RAG)** based framework. The process starts with a **user query** which in this case is the body of a business e-mail from the Enron Email Dataset, which is processed by **Information Retrieval (IR)** models (BM25 or FAISS) to retrieve relevant passages from the pre-existing curated **Knowledge Base (KB)**. The top- $K$  retrieved passages, or the raw query in the case of zero-shot classification, are used as input to a **Large Language Model (GPT-4o)** for generating classification labels and detailed natural language explanations.

#### A. Knowledge Base (KB) Construction

The considered KB was created by curating a high-quality repository of privacy-related content, specifically scientific articles on the topic. We retrieved 100 documents using two main strategies.

- The first strategy involved a set of open-access versions of academic papers sourced through manual selection from top-level privacy and security conferences, ensuring the inclusion of authoritative research and recent developments in the field. The selected conferences include the *Privacy Enhancing Technologies Symposium (PETS)*, *IEEE Symposium on Security and Privacy*, *IEEE European Symposium on Security and Privacy (EuroS&P)*, *ACM Conference on Computer and Communications Security (CCS)*, and the *Workshop on Privacy in the Electronic Society (WPES)*;
- The second strategy employed an automated retrieval process, followed by manual verification, using the Semantic Scholar APIs.<sup>2</sup> This search targeted papers specifically addressing privacy and data sensitivity within the contexts of *Natural Language Processing (NLP)* and *ML*, using the following query:

```
(`sensitive data` OR `sensitive
information` OR `confidential
data` OR `confidential
information` OR `personal data`
OR `personally identifiable
information (PII)`) AND (`LLM`
OR `NLP` OR `RAG`)
```

<sup>2</sup><https://www.semanticscholar.org/product/api>

The selection of 100 papers for the KB was chosen to provide a sufficient coverage or retrieving privacy-sensitive information using the RAG framework. This number represents a balance between offering enough diversity in content for effective retrieval and maintaining a manageable manual quality control. The KB was curated in order to be intended as a general purpose and adaptable to various domains, despite not fully explored in this work. To concentrate on the core technical aspects of privacy risks, we excluded general regulations (e.g., GDPR, CCPA) from the KB. These regulations, while essential for compliance, often introduce broader legal and procedural considerations that can obscure the specific technical nuances relevant to our analysis. Finally, after the automated retrieval, each document was manually reviewed to confirm its relevance to privacy concerns, ensuring that only contextually relevant papers were included.

#### B. Dataset Creation and Annotation

The *Enron Email Dataset* [19] is a publicly available collection of approximately 500,000 emails from 150 executives and employees of Enron Corporation, covering the period from 2000 to 2002.<sup>3</sup> Released during the company's bankruptcy investigation, it has become a key resource for research in *ML*, *Natural Language Processing (NLP)*, and *Social Network Analysis (SNA)*. The dataset includes both metadata (i.e., *sender*, *recipient*, *timestamp*) and message content, with some versions containing attachments.

The dataset was then filtered by selecting 3,000 random emails. After a data cleaning process that removed overly long and overly short email bodies (excluding those under 50 characters and over 2,000 characters), we obtained a dataset

<sup>3</sup><https://www.cs.cmu.edu/~enron/>

of 1,223 instances. For the dataset annotation, the following process was employed. We conducted a human validation assessment to evaluate GPT-4o’s ability to detect whether a text contains private or sensitive information. For this purpose, we conducted an initial annotation of 120 instances from the Enron Mail Dataset, involving a legal expert and GPT-4o using a few-shot approach. The model was provided with 4 human-annotated examples that were not part of the initially selected 120.

We measured the inter-annotator agreement and obtained a Cohen’s kappa score of 0.8, which is in the range 0.6–0.8 that indicates a “substantial agreement” [20]. This score demonstrates GPT-4o’s ability to correctly understand the task. As a result, we relied on GPT-4o to annotate the 1,103 remaining entities of the dataset.

### C. Information Retrieval (IR) Models

For the retrieval process involved in RAG we employed two state-of-the-art IR models for both *sparse* (BM25) and *dense* (FAISS) retrieval.

- The BM25 model was employed as implemented in the `LlamaIndex` library,<sup>4</sup> which provides a robust and efficient sparse retrieval mechanism based on the *Term Frequency–Inverse Document Frequency* (TF–IDF) text representation model. This method works particularly well in scenarios where exact term matching is crucial, making it suitable for retrieving privacy-related passages from the KB;
- FAISS was employed as implemented in the `faiss` python library,<sup>5</sup> and represents the documents as a dense vector using the `msmarco-distilbert-base-v3` sentence model.<sup>6</sup> This method enables semantic matching, allowing the system to retrieve passages that are contextually similar to the query, even if exact keywords are not present. For our aim, this is particularly important in cases where privacy violations are implied rather than explicitly stated.

### D. The Considered Prompt

The *prompt* used to guide the LLM in the RAG framework in the classification process ensures a structured approach to identify privacy violations. It consists of a preamble followed by the text to be analyzed and the relevant passages retrieved during the retrieval phase. It then requires providing a classification label and a clear explanation to justify this assignment. It is defined as illustrated in Figure 2.

## IV. EXPERIMENTAL SETUP

The experiments have been conducted using the OpenAI’s APIs.<sup>7</sup> We set a low temperature of 0.2 and limited the output to 512 tokens. The low temperature was chosen to minimize

<sup>4</sup>[https://docs.llamaindex.ai/en/stable/examples/retrievers/bm25\\_retriever/](https://docs.llamaindex.ai/en/stable/examples/retrievers/bm25_retriever/)

<sup>5</sup><https://pypi.org/project/faiss/>

<sup>6</sup><https://huggingface.co/sentence-transformers/msmarco-distilbert-base-v3>

<sup>7</sup><https://openai.com/blog/openai-api>

*Your task is to analyze a text to determine whether it exposes private or sensitive information, and provide an explanation of your answer. The text to be analyzed is the following: { input text }.*

*Base your answer on the following passages: { relevant passages }. (RAG-ONLY)*

*Format your answer by providing both a label and a clear and thorough explanation, as follows:*

*Label: it should be “violation detected” or “no violation detected”.*

*Explanation: it should be a complete and well-reasoned analysis, referencing the provided documents to justify why the text shows violations or not.*

Fig. 2. Prompt adopted in the proposed solution.

the “creativity” of GPT-4o and ensure the most consistent results as possible for the classification tasks.

Below, we present the evaluation of the different retrieval and classification approaches for identifying sensitive or private content violations within email bodies in the Enron Email Dataset introduced before. The primary objective is to assess how these various retrieval models impact the classification performance of the generative model, specifically GPT-4o compared to a zero-shot approach.

### A. Methods

We conducted distinct experiments using GPT-4o as the base model of the architecture:

- First, we performed a *zero-shot classification* without introducing any retrieved information from the KB as a baseline method;
- Then, we experimented *RAG-based classification* with both BM25 and FAISS on different setups of the KB: *Abstracts only* (A), *Introductions only* (I), and finally on *both* (A + I).

Table I shows the results of the zero-shot baseline compared to the three setups in terms of *Precision*, *Recall*, and *F1-Score*.

In general, it can be observed that the RAG-based solution using sparse retrieval with BM25 outperforms other models in each KB setup, achieving the highest Recall and F1-Scores. RAG based on dense retrieval shows significantly lower Recall, suggesting it may retrieve documents that are semantically similar but less relevant for specific privacy-sensitive keywords, thereby reducing its overall effectiveness. The zero-shot solution achieves the highest Precision across all setups, indicating that it is more conservative in identifying “violation detected” cases. However, it still misses many true positives, leading to a significant drop in both Recall and F1-Scores. These results show that, overall, the abstract-only setup yields the best performance, suggesting that the structure of abstracts provides a more focused source

TABLE I  
RESULTS FOR ZERO-SHOT CLASSIFICATION AND RAG-BASED CLASSIFICATION WITH BM25 AND FAISS ACROSS DIFFERENT KB SETUPS.

Setup	Method	Precision	Recall	F1-Score
<i>zero-shot</i>	-	<b>0.8812</b>	0.6987	0.7283
<i>Abstracts (A)</i>	<b>BM25</b>	0.8763	<b>0.7548</b>	<b>0.7795</b>
	FAISS	0.7913	0.5988	0.6940
<i>Introductions (I)</i>	BM25	0.8675	<b>0.6148</b>	<b>0.7621</b>
	FAISS	0.8256	0.6115	0.7482
A + I	BM25	0.8696	<b>0.7334</b>	<b>0.7612</b>
	FAISS	0.8165	0.6438	0.7393

of privacy-relevant information. On the other hand, the use of introductions often contains a mix of background information, problem statements, and research motivations, which can introduce content unrelated to specific privacy concerns, especially when using dense retrieval models in the RAG framework.

### B. Explainability

Along with the privacy violation classification, we instructed the model to generate related explanations – some examples are illustrated in Table II – which revealed that the zero-shot approach primarily focused on explicit identifiers and business context but was limited in detecting broader policy or regulation violations. It classified emails as “violation detected” only if they explicitly contained direct identifiers, while almost everything else was marked as “no violation detected”, consistent with its Precision and Recall metrics. The RAG method based on sparse retrieval using BM25 seems to adopt a more general approach, considering certain business domain policies loosely and occasionally referencing external regulations, such as GDPR. In contrast, the RAG method based on dense retrieval using FAISS shows limitations in retrieving broader privacy-related information, often focusing on business-related content. This limitation leads to missed detections of key privacy violations, as demonstrated by the results shown earlier.

## V. FURTHER REMARKS

In this work, we propose a framework that demonstrates the potential of RAG systems in automating privacy-sensitive content detection and providing clear contextual explanations in natural language. Although the results are promising with the tested SOTA GPT-4o, we acknowledge that to fully understand the framework’s capabilities and limitations, broader testing with other OpenAI models and open-source alternatives is essential. Moreover, the KB and retrieved passages could benefit from further refinement (e.g., embeddings augmentation for FAISS and specific word selection or filtering for BM25) to enhance the precision and contextual relevance of the information provided to the language model.

These initial experimental results indicate a high level of agreement between automated and human detection of

TABLE II  
EXAMPLES OF SOME ANSWERS OBTAINED ACROSS DIFFERENT METHODS, I.E., NON-RAG-BASED ZERO-SHOT AND RAG-BASED (WITH BM25 AND FAISS IR MODELS).

Method	Answer
zero-shot	<b>Email:</b> When you guys do loss return deals like this one with PGE, could you include a note in the comments section regarding the zero price and “no confirm” status? ... <b>Label:</b> <i>No violation detected.</i> <b>Explanation:</b> <i>The content provided pertains to general business information and does not include identifiable personal data, financial information, or sensitive corporate strategies that would typically be classified as private or confidential.</i>
	<b>Email:</b> ... availability of your current Natural Gas Intelligence Newsletter(s). Please use your username of “kward” and your password to access ... <b>Label:</b> <i>violation detected.</i> <b>Explanation:</b> <i>The text contains sensitive information, specifically the username “kward” and a password for accessing subscription-based services provided by Natural Gas Intelligence. Sharing login credentials publicly or through unsecured channels can lead to unauthorized access to personal accounts and services, which is a violation of privacy and security protocols.</i>
BM25	<b>Email:</b> Wes, As we discussed on Friday, I think that there is merit to your arguments regarding the release in prudence in North America. However, this may alter the way we articulate the relative liquidity ... <b>Label:</b> <i>No violation detected.</i> <b>Explanation:</b> <i>The text provided does not appear to violate any specific privacy policies based on the information given. The content of the text is a professional communication ...</i>
	<b>Email:</b> ... Could you please give me a Yes, No, Maybe decision on the following candidates interviewed: Bruce Kamich, Philip Roan, Jerzy Jarosz, Bruce James, Oleg Bondar. Thanks, Toni ... <b>Label:</b> <i>Violation detected.</i> <b>Explanation:</b> <i>According to privacy policies and data protection regulations, such as the General Data Protection Regulation (GDPR) in the EU or the California Consumer Privacy Act (CCPA) in the US ... the text contains sensitive information related to the recruitment process at Enron North America Corp. It includes the names of several individuals who were candidates for a position, as well as internal communications discussing their evaluations. This information can be considered Personally Identifiable Information (PII) because it identifies specific individuals in the context of a job application process. ...</i>
FAISS	<b>Email:</b> Norma, The new position created it that for a Staff Specialist, who would be responsible for technical duties including evaluation and quality control of weather forecasting procedures, design and preparation of the research Intelligence newsletter ... <b>Label:</b> <i>No violation detected.</i> <b>Explanation:</b> <i>The text provided is an internal communication regarding the creation of a new position within a research group and the intention to fill this position with an existing employee, William (Sam) Smith. The content of the message primarily pertains to organizational and operational details, such as the responsibilities associated with the new position and the proposed retroactive start date. From a privacy policy compliance perspective, the text does not appear to violate any specific privacy rules. ...</i>
	<b>Email:</b> Thursday’s (4/6/00) conference call regarding the weekly Japan update will now be held at 6:00 p.m. Houston time and 8:00 a.m. Tokyo time due to Daylight Savings time change. Please forward your updates to Jope Hirl and copy me. The call will be held in Conference room 21C2 for Houston participants. Thanks, Eddie Robles Domestic: 1-800-991-9019 Intern’l 847-619-8039 Code: 5764488 <b>Label:</b> <i>Violation detected.</i> <b>Explanation:</b> <i>The text provided contains sensitive information that could potentially violate privacy policies. The following points highlight why the text is considered sensitive: 1. **Personal Identifiable Information (PII):** ... 2. **Contact Information:** ... 3. **Conference Call Details:** ... 4. **Time and Location Information:** ...</i>

privacy-sensitive data (Cohen’s kappa reaching a score of 0.8). However, our findings reveal considerations that warrant further attention in subsequent research. This need is particularly highlighted during the dataset annotation process, which uncovered significant inconsistencies in the automated identification of corporate telecommunications identifiers. For instance, the model classified corporate telephone numbers as sensitive data while excluding corporate email addresses. These findings underscore the need to develop more sophisticated classification systems capable of distinguishing various categories of “privacy-sensitive data” in accordance with applicable regulatory frameworks, including judicial and administrative interpretations.

Lastly, the architecture’s performance has only been evaluated within the context of a single dataset. Extending the evaluation to other diverse datasets across various domains is necessary to assess its generalizability and adaptability to different contexts.

## VI. CONCLUSIONS AND FURTHER RESEARCH

In the context of automated detection of private and sensitive information violations in unstructured data, our study demonstrated the potential of a Retrieval-Augmented Generation (RAG) framework that integrated BM25 and FAISS retrieval models to enhance the detection capabilities of Large Language Models (LLMs). By leveraging a structured Knowledge Base (KB) composed of scientific articles on data privacy and confidentiality, the proposed system showed notable improvements in identifying privacy-sensitive information while generating human-readable explanations in natural language. As an initial investigation, we limited our evaluation to the state-of-the-art GPT-4o model to validate the feasibility of our approach. To ensure a controlled and rigorous assessment, we focused on a single dataset within a specific domain, which allowed us to prioritize the development of a robust and well-structured pipeline architecture. Experimental results indicated that incorporating a curated KB significantly enhanced the performance of GPT-4o, reinforcing the potential of retrieval-augmented methods in improving privacy-related text analysis.

For future work, we aim to conduct a comprehensive evaluation and comparison of different models, considering various sizes, architectures, and open-source alternatives to assess their effectiveness and applicability. Additionally, we plan to extend dataset evaluations by incorporating a broader range of domains and languages to enhance the generalizability of our findings. Further improvements to the KB will be explored, including more advanced filtering techniques and refined processing of extracted segments to ensure higher-quality Information Retrieval. Moreover, we will perform in-depth qualitative analyses of the model’s explanations by systematically comparing them with those provided by human experts. These comparisons will be further validated through user studies and real-world case studies, involving end-users to assess the practical utility, trustworthiness, and interpretability of the generated explanations.

## ACKNOWLEDGMENTS

This work was partly supported by: the EC under grant GLACIATION (101070141), the EU-Next Generation EU, Mission 4, Component 2, CUP: D53D23008480001 (KURAMI: Knowledge-based, explainable User empowerment in Releasing private data and Assessing Misinformation in online environments),<sup>8</sup> the EU-Next Generation EU, Mission 4, Component 2, CUP: J33C22002830006 (ExtrAI), and by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the Italian MUR. Neither the European Union nor the Italian MUR can be held responsible for them.

## REFERENCES

- [1] C. Dwork, A. D. Smith, T. Steinke, and J. Ullman, “Exposed! a survey of attacks on private data,” *Annual Review of Statistics and Its Application*, vol. 4, pp. 61–84, 2017.
- [2] S. S. Bakare, A. O. Adeniyi, C. U. Akpuokwe, and N. E. Eneh, “Data privacy laws and compliance: A comparative review of the EU GDPR and USA regulations,” *Computer Science & IT Research Journal*, 2024.
- [3] H. Attaullah, A. Anjum, T. Kanwal, S. U. R. Malik, A. Asheralieva, H. Malik, A. Zoha, K. Arshad, and M. A. Imran, “F-classify: Fuzzy rule based classification method for privacy preservation of multiple sensitive attributes,” *Sensors (Basel, Switzerland)*, vol. 21, 2021.
- [4] G. Livraga, A. Olzjojevs, and M. Viviani, “Unveiling the privacy risk: A trade-off between user behavior and information propagation in social media,” in *International Conference on Complex Networks and Their Applications*, pp. 277–290, Springer, 2023.
- [5] L. Cassani, G. Livraga, and M. Viviani, “Assessing document sanitization for controlled information release and retrieval in data marketplaces,” in *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 88–99, Springer, 2024.
- [6] J.-S. Park, G.-W. Kim, and D.-H. Lee, “Sensitive data identification in structured data through GenNER model based on text generation and ner,” *Proceedings of the 2020 International Conference on Computing, Networks and Internet of Things*, 2020.
- [7] S. Anand, M. Shukla, and S. Lodha, “Detecting sensitive information from unstructured text in a data-constrained environment,” *2023 15th International Conference on COMMunication Systems & NETWORKS (COMSNETS)*, pp. 159–164, 2023.
- [8] J. Neerbeky, I. Assentz, and P. Dolog, “Taboo: Detecting unstructured sensitive information using recursive neural networks,” in *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pp. 1399–1400, 2017.
- [9] Y. Hu, B. Guo, C. Dai, Q. Zheng, F. Weng, and Z. Li, “A fine-grained multi-label privacy detection model for unstructured data based on bert pre-training,” *2022 IEEE 24th Int Conf on High Performance Computing & Communications; 8th Int Conf on Data Science & Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pp. 755–760, 2022.
- [10] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, and J. rong Wen, “A survey of large language models,” *ArXiv*, vol. abs/2303.18223, 2023.
- [11] N. Thalpage, “Unlocking the black box: Explainable artificial intelligence (xai) for trust and transparency in ai systems,” *Journal of Digital Art & Humanities*, 2023.
- [12] P. Atanasova, J. Simonsen, C. Lioma, and I. Augenstein, “A diagnostic study of explainability techniques for text classification,” *ArXiv*, vol. abs/2009.13295, 2020.
- [13] W. Yang, Y. Wei, H. Wei, Y. Chen, G. Huang, X. Li, R. Li, N. Yao, X. Wang, X. Gu, M. Amin, and B. Kang, “Survey on explainable ai: From approaches, limitations and applications aspects,” *Human-Centric Intelligent Systems*, vol. 3, 08 2023.

<sup>8</sup><https://kurami.disco.unimib.it/>

- [14] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of Retrieval Augmented Generation (RAG) models for open domain question answering," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, 2022.
- [15] Z. Yue, H. Zeng, Y. Lu, L. Shang, Y. Zhang, and D. Wang, "Evidence-driven retrieval augmented response generation for online misinformation," in *North American Chapter of the Association for Computational Linguistics*, 2024.
- [16] R. Upadhyay and M. Viviani, "Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy," *Discover Computing*, 2025. To appear.
- [17] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The Faiss library," 2024.
- [18] D. Boumber, B. E. Tuck, R. M. Verma, and F. Z. Qachfar, "Llms for explainable few-shot deception detection," in *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics*, pp. 37–47, 2024.
- [19] E. Corp and W. W. Cohen, "Enron email dataset." <https://www.loc.gov/item/2018487913/>, 2015. [Philadelphia, PA: William W. Cohen, MLD, CMU] [Software, E-Resource].
- [20] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.