# Exploring the Application of Large Language Models in Detecting and Protecting Personally Identifiable Information in Archival Data: A Comprehensive Study*

Jianliang Yang
*School of Information Resource Management*
*Renmin University of China*
Beijing, China
rucyjl1991@ruc.edu.cn

Xiya Zhang
*School of Information Resource Management*
*Renmin University of China*
Beijing, China
zhangxiya@ruc.edu.cn

Kai Liang
*Digital Archives Management Office*
*Hangzhou Archives*
Zhejiang, China
lk_114@163.com

Yuenan Liu*
*School of Information Resource Management*
*Renmin University of China*
Beijing, China
liuyuenan@ruc.edu.cn

***Abstract*—This comprehensive study investigates the application of Large Language Models (LLMs) for detecting and protecting Personally Identifiable Information (PII) in archival data, a pressing concern for archives under the mandate to increase public access while safeguarding personal privacy. The paper juxtaposes traditional supervised learning methods against LLMs' unsupervised capabilities in PII detection, unveiling LLMs as viable alternatives capable of achieving satisfactory performance levels without the need for extensive training datasets. Through empirical analysis, the study validates the feasibility of LLMs in identifying sensitive information within large volumes of archival material. The findings highlight LLMs' significant interpretability, providing understandable rationale behind PII identification—a feature that not only enhances trust in AI applications but also aids archival staff in the review process. This research contributes novel insights into the intersection of AI and archival science, presenting LLMs as powerful tools for addressing the twin challenges of data accessibility and privacy.**

***Keywords—Personally Identifiable Information; Large Language Models; Archival Data; AI Explainability***

## I. INTRODUCTION

The identification of Personally Identifiable Information (PII) within archival data is both a crucial and challenging task. In China, numerous archives are under immense pressure to open their collections to the public, yet they are hindered by a lack of adequate resources to address privacy concerns adequately. Researchers and practitioners alike have proposed the use of artificial intelligence (AI) to detect PII within these vast stores of historical data [1, 2, 3]. However, this approach is commonly beset with obstacles, including a dearth of training data and difficulties in achieving interpretable results.

The advent of Large Language Models (LLMs) has illuminated an alternative path. LLMs hold the promise of conducting unsupervised detection of PII, which could not only sidestep the issue of scarce training materials but also enhance the interpretability of the detection process. The interpretability aspect is particularly compelling, as it aligns with the increasing demand for transparent AI systems that stakeholders can trust and understand.

This study undertakes an empirical exploration into the use of LLMs for the detection of PII. By harnessing the capabilities of these advanced models, the research seeks to uncover the extent to which LLMs can be employed effectively in this domain. It aims to contribute to the body of knowledge by demonstrating how LLMs can be leveraged to address the pressing need for efficient and reliable PII detection in archival documents, thereby facilitating the safe disclosure of historical information in a manner that respects individual privacy and adheres to regulatory standards.

## II. RELATED WORK

### A. Utilizing AI for Detecting PII

PII is ubiquitously embedded within documents or databases resulting from interactions between users and public institutions (e.g., governments, enterprises). Particularly for entities with an openness mandate, like archival institutions, striking a balance between open access to information and the protection of personal data remains a formidable challenge for the forthcoming future. This is because archival materials require review by authoritative bodies prior to their release, to ensure that they do not contain state secrets or other non-disclosable information, including PII. Given the sheer volume of archival materials awaiting review, relying on manual inspection is insufficient to meet the public's demand for transparency. With so large-scale "Dark" archives, AI

* This work was supported by the National Social Science Fund of China under No. 21CTQ030. The views expressed in this work represent the opinions of the authors and may not necessarily be shared or endorsed by their respective employers and/or sponsors.

* The Corresponding Author.

technology has been imported to identify PII within archives [4].

In early technology-assisted solutions, regular expression [5], Named Entity Recognition (NER) [6], and topic modeling [7] are frequently utilized to identify PII in archival materials. These methods can precisely identify fields containing specific types of PII that archivists have already realized and intentionally controlled. However, they fall short when dealing with complex semantic texts, and merely aid archivists in reducing the scope of their review to some extent . The emergence of supervised machine learning has equipped computer with the capacity to comprehend context, facilitating an in-depth semantic recognition of PII with higher identification rate. And T. Hutchinson [8] corroborated this through a comparative study wherein his team had previously employed topic modeling to detect PII within HR-related documents. Nevertheless, supervised machine learning relies on the training of fully labelled datasets, which means a significant amount of manual annotation in the preliminary stage, and this approach can only be repurposed within a specific domain [9].

As the rise of deep learning techniques, models like Convolutional Neural Networks (CNN), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM) have achieved excellent performance in large-scale text classification without labels [10]. da Silva [11] conducted an in-depth exploration of the application of CNN in NER tasks for the PII detection. Meanwhile, Poornima Kulkarni and Cauvery N K [12] developed a clustering-based PII Model (C-PPIM) based on NLP and Byte-mLSTM, which can be used to automatically detect PII in unstructured text corpus. However, due to the “black box” nature of AI, the professional responsibility and ethics of archivists, algorithmic bias, and other complex factors [13], the progression of AI tools in identification applications within archival institutions has been slow.

### B. *Capabilities of Large Language Models*

LLMs are a type of natural language processing models characterized by the vast scale of parameters, yet their essence is to estimate the probability distribution over text [14]. The remarkable capabilities of an LLM offer archivists the potential to identify PII within archival materials with enhanced efficiency and quality. This can be broken down into the following aspects: Firstly, LLMs are Few-Shot Learners [15] or Zero-Shot Reasoners [16], which means the pre-training of LLMs requires only a minimal number of samples or no samples at all. LLMs can extract complex features from vast amounts of data, possessing robust capabilities in content understanding and generation, with outstanding generalization performance [17]. Secondly , LLMs possess a robust capability for context comprehension; which can discern hidden patterns within massive corpus and generate accurate predictions[17]. This capability is crucial for the identification and protection of PII, since information that can directly or indirectly identify an individual's identity is referred to as PII. On the one hand, PII may be dispersed amidst extensive textual data, making it challenging to detect through traditional rule-based matching. On the other hand, anonymized personal information might still be re-identified due to associations with other pieces of information [18]. Hence, the powerful contextual understanding capability of LLMs enables a semantic excavation, association, and identification of PII, providing stronger protection for PII in archival materials that should not be disclosed. Thirdly, LLMs can simulate human-like chain-of-thought reasoning. Through chain-of-thought prompting, these models can accomplish intricate tasks step-by-step and provide logical explanations [19], substantially enhancing the explainability of models. “It is difficult to understand why the machine makes the decisions it makes.” [20] Similar concerns related to the “black box” nature of AI can be alleviated to some extent.

Overall, LLMs can detect numerous PII risks that traditional methodologies might overlook and provide reasoning to assist archivists in their review. Furthermore, LLMs can continually enhance their identification capabilities through adaptive learning, which has significant potential for prudent identification of PII in archival materials. Thus, this study delineates the implementation of LLMs in PII detection.

## III. Problem and Methods

### A. *Challenges of PII within Hangzhou Archives*

Following directives from regulatory authorities, HangZhou Archives faces the unprecedented challenge of making approximately 2 million volumes of archives accessible to the public within the next five years. The magnitude of this disclosure imposes a significant burden on the vetting process. In response, HangZhou Archives has developed a detailed review system that encompasses multiple levels of access: public release, governmental intranet availability, organizational internal access, and non-disclosure. Additionally, they have established a rigorous procedure for archive release and key guidelines for reviewing archives. One of these clear guidelines is that archives containing personal privacy information cannot be released to the public.

A central task in the review process is to sift through the vast collection and identify archives containing personal data. This intricate and labor-intensive task is further challenged by the current staffing levels at HangZhou Archives. Complications arise when the review process involves collaboration across multiple units, with ambiguities in procedural roles amplifying the challenges. On one side, HangZhou Archives grapples with this massive vetting workload; on the other, it contends with limited resources and a lack of efficient methods to carry out the task. Such constraints have nudged HangZhou Archives towards considering the integration of AI technologies to facilitate the review process.

While several AI solutions exist in the market, they often face the challenge of insufficient training data. HangZhou Archives might provide definitions for personal and private content but lacks well-curated datasets for AI training. AI service providers, thus, are tasked with the annotation of data that contains personal information, crucial for training their models. This annotation process can be resource-intensive, demanding considerable time and effort. Given that archives from various sources remain sealed until vetted for public release, and may contain intricate, confidential information, this significantly raises the bar for annotation tasks. Additionally, the outcomes from these AI solutions often lack sufficient explainability, causing review personnel to approach them with caution.

Existing AI approaches typically frame the problem of identifying personal privacy in archives as an artificial intelligence classification model. Let the dataset be denoted as:

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \qquad (1)$$

where each $x_i$ represents the textual information of an archive, and each $y_i$ indicates whether the archive pertains to personal privacy (e.g., 1 signifies privacy involvement, 0 signifies no involvement). The objective is to learn a function:

$$f: \mathcal{X} \rightarrow \mathcal{Y} \qquad (2)$$

where $\mathcal{X}$ represents the set of all potential textual data, and $\mathcal{Y} = \{0,1\}$.

To identify the optimal function $f^*$, we can employ a particular loss function $\mathcal{L}(y, f(x))$ to denote the discrepancy between the actual category $y$ and the predicted category $f(x)$ The goal is to minimize the average loss across the entire dataset:

$$f^* = \arg \min_{f} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\,(y_i, f(x_i)) \qquad (3)$$

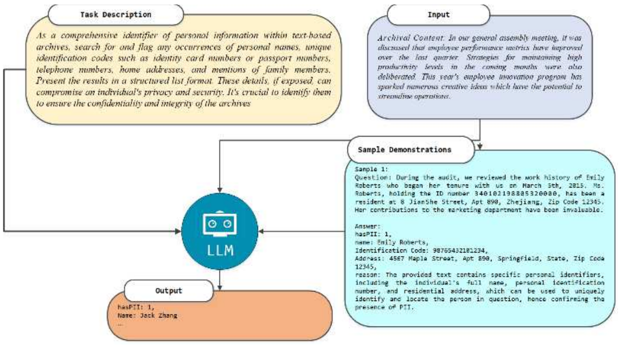
In practical machine learning settings, the function f is generally represented by certain models (such as neural networks or decision trees) that possess specific parameters, which are learned using optimization algorithms.

### *B. Unsupervised Approaches Based on LLM*

In the rapidly evolving field of artificial intelligence, the advent of LLMs such as GPT-series has paved the way for a myriad of innovations. One of the most intriguing aspects of LLMs is their emergent behaviors, which bestow them with impressive generalization capabilities. Such behaviors, often unexpected and not explicitly programmed, result from the massive amount of data these models have been trained on.

A striking manifestation of this generalization ability is seen in the way LLMs handle diverse tasks. Recent studies have reported the application of models like ChatGPT for tasks such as sentiment analysis [21] and information extraction [22]. These tasks traditionally required specific architectures or fine-tuning on labeled datasets. However, with ChatGPT and its counterparts, researchers found that the model, straight out of the box, was adept at making nuanced decisions and inferences, drawing on its vast training data. This capability demonstrates the emergent phenomena inherent in LLMs and underlines the inherent versatility and multitasking prowess these models possess.

One breakthrough that plays a crucial role in harnessing the power of LLMs is the technique of "prompting". Prompting, in essence, refers to the art of querying or instructing the model in a way that extracts the desired behavior or information [23] . Instead of going through the time-consuming process of training or fine-tuning models on specific datasets, one can skillfully design prompts that guide the LLM to produce the required output. This drastically reduces computational resources and time required, making applications more scalable and adaptable. Designing effective prompts, however, is an art in itself. It often requires a deep understanding of the model's behavior and its response mechanisms. Crafting a prompt that can accurately and reliably extract information or guide the model's behavior is essential for practical applications. Given that LLMs do not come with a fixed user manual and their internal workings are largely opaque, finding the right prompt often involves a mix of expertise, experimentation, and intuition.



*As a comprehensive identifier of personol information within text-hosed archives, search for ond flag any occurrences of personal names, unique identification codes such as identity card numbers or passport numbers, telephone numbers, home addresses, and mentions of family members. Present the results in a structured list format. These details, if exposed, can compromise an individual's privacy and security. It's crucial to identify them to ensure the confidentiality and integrity of the archives*

*Archival Content: In our general assembly meeting, it was discussed that employee performance metrics have improved over the last quarter. Strategies for maintaining high productivity levels in the coming months were also deliberated. This year's employee innovation program has sparked numerous creative ideas which have the potential to streamline operations.*

```
Sample 1:
Question: During the audit, we reviewed the work history of Emily
Roberts who began her tenure with us on March 5th, 2015. Ms.
Roberts, holding the ID number 340102198805320080, has been a
resident at 8 JianShe Street, Apt 890, Zhejiang, Zip Code 12345.
Her contributions to the marketing department have been invaluable.

Answer:
hasPII: 1,
name: Emily Roberts,
Identification Code: 98765432101234,
Address: 4567 Maple Street, Apt 890, Springfield, State, Zip Code
12345,
reason: The provided text contains specific personal identifiers,
including the individual's full name, personal identification
number, and residential address, which can be used to uniquely
identify and locate the person in question, hence confirming the
presence of PII.
```

```
hasPII: 1,
Name: Jack Zhang
...
```

Fig.1. The Model Architecture for Detecting PII Using an Unsupervised LLM

Now, considering the vast capabilities of LLMs and the power of prompting, it becomes conceivable to design an unsupervised approach to detect personal information within archives. Archives, often voluminous and with diverse content, present a unique challenge. Manually sifting through these archives to identify and redact personal information is labor-intensive and fraught with potential oversight. By leveraging LLMs, a potential methodology could involve designing prompts that are inherently sensitive to personal information cues. The model architecture is shown in Figure 1, for instance, prompts could be crafted to make the model responsive to patterns typically associated with personal details like names, addresses, dates, and other identifiers. When exposed to a vast archive, the LLM could then flag potential segments of text that match these patterns. Given the model's vast training data, it's likely already familiar with a plethora of formats and representations of personal information, making it a formidable tool for this task.

In summation, the synthesis of LLMs' emergent faculties with the refined art of prompting augurs a transformative epoch in AI applications. The potential of harnessing these methodologies for the unsupervised discernment of personal data within archives underscores the versatility and dynamism of contemporary AI paradigms. As this academic domain burgeons, it is poised to unveil a cornucopia of innovative applications, heralding a renaissance in traditional processes and paradigms.

### *C. Constructing Prompts for PII*

In the realm of archiving, managing and protecting personal information is of paramount importance. With the rise of LLMs, we now have sophisticated tools to assist in identifying and safeguarding such personal details in archives. Yet, the effectiveness of these models largely depends on how we frame our instructions to them, which, in AI jargon, is termed as "prompting". The challenge and opportunity here revolve around the construction of prompts that precisely target the desired personal information.

*1) Understanding the Landscape of PII in Archives:* To harness the potential of an LLM effectively, it is essential to first understand the nature and scope of personal information commonly found in archives. Through meticulous analysis, it is discerned that archives frequently contain the following categories of personal information:

- **Personal Names:** The most straightforward and yet pivotal form of personal identification.
- **Personal Identification Codes:** These encompass a range of unique identifiers such as National Identity

Card numbers, Driver's License numbers, and Passport numbers.

- **Telephone Numbers:** Often linked directly to an individual or household.
- **Home Addresses:** Specific locations that can give away an individual's residence or historical places of stay.
- **Family Members:** Information that provides details on an individual's immediate or extended family, possibly exposing relationships and family structures.
- **Other information combined that can identify an individual.**

The primary objective of LLMs, in this context, is to meticulously search and flag archives that contain any of these specified categories of personal information.

*2) The CRSRF Prompts Framework for Detecting PII:* Harnessing the capability of an LLM to identify personal information in archives requires a methodological approach to crafting prompts. This involves understanding not only what we want the model to do but also ensuring the model comprehends the depth, implications, and importance of the task. Inspired by the CRISP prompt engineering method [24], a framework named 'CRSRF' for constructing prompts is bifurcated into three main components:

- **Capacity and Role:** This section establishes the LLM's task. It informs the model of its role as a detector and protector of personal information within archives. The prompt might start with phrasing like, *"As a comprehensive identifier of personal information within text-based archives..."*
- **Statement:** This defines the exact task at hand. It explicitly states the kind of personal information the LLM should look for. Given our analysis of personal information in archives, the statement can be framed as: *"Search for and flag any occurrences of personal names, unique identification codes such as identity card numbers or passport numbers, telephone numbers, home addresses, and mentions of family members..."*
- **Reason:** This provides a context or justification for the task, underscoring the importance of safeguarding personal data. Including a reason can enhance the model's alignment with the intent of the task. A sample reason might be: *"These details, if exposed, can compromise an individual's privacy and security. It's crucial to identify them to ensure the confidentiality and integrity of the archived documents."*
- **Format:** This section specifies the desired output format for the information extracted by the LLM. Given the structured nature of data, formats such as lists are preferable. For instance, *"Present the identified personal information in a list format, with categories such as 'Name', 'Identification Code', 'Telephone Number', 'Address', and 'Family Members' as keys."*

A well-constructed prompt following this framework could look something like:

*" As a comprehensive identifier of personal information within text-based archives, search for and flag any occurrences of personal names, unique identification codes such as identity card numbers or passport numbers, telephone numbers, home addresses, and mentions of family members. Present the results in a structured list format. These details, if exposed, can compromise an individual's privacy and security. It's crucial to identify them to ensure the confidentiality and integrity of the archives "*

The significance of accurately identifying personal information in archives can't be understated. With LLMs at our disposal, we are equipped with a powerful tool to undertake this task. Yet, the key to harnessing this power lies in the art of prompt construction. By understanding the nature of personal information in archives, considering the desired output format, and employing a structured framework for prompt crafting, we can effectively guide an LLM in safeguarding invaluable personal data within archival settings. As we move forward, it becomes imperative to refine and perfect these prompts, making them more nuanced and aligned with the evolving nature of personal data.

*3) Enhancing LLM with Contextual Samples in the CRSRF Prompts Framework:* In addition to crafting precise prompts that clearly articulate the task at hand, it is equally important to provide contextually relevant examples within the prompts to orient the LLM towards more accurate performance [25]. This method is akin to providing 'samples' that serve as a reference for the kind of information the model needs to identify and flag. Such examples can significantly enhance the model's ability to discern between different types of PII and non-PII information.

For instance, a refined prompt may include:

*"As you meticulously search for personal information within the given archive, consider these examples as a benchmark for you to answer.*

*Question: During the audit, we reviewed the work history of Emily Roberts who began her tenure with us on March 5th, 2015. Ms. Roberts, holding the ID number 340102198805320000, has been a resident at 8 JianShe Street, Apt 890, Zhejiang, Zip Code 12345. Her contributions to the marketing department have been invaluable.*

*Answer:*

*hasPII: 1,*

*name: Emily Roberts,*

*Identification Code: 340102198805320000,*

*Address: 8 JianShe Street, Apt 890, Zhejiang, Zip Code 12345,*

*reason: The provided text contains specific personal identifiers, including the individual's full name, personal identification number, and residential address, which can be used to uniquely identify and locate the person in question, hence confirming the presence of PII."*

By embedding such specific instances in the prompts, we are not only defining the boundaries of PII for the LLM but also equipping it with a clearer understanding of the nuances involved in what is considered private and sensitive. The integration of these 'samples' into the CRSRF Prompts Framework is an essential step in advancing the model's capability to make informed and context-aware decisions. It ensures that the model not only follows the directive to locate PII but also develops a more sophisticated grasp of the PII elements it is tasked to detect. Consequently, this approach facilitates the model's learning process, enabling it to become increasingly proficient at identifying a wider array of PII with greater precision, thereby fortifying the confidentiality and integrity of archival materials.

## IV. Data and Evaluation

Within the complex matrix of information processing, the efficacy of any system is intrinsically linked to the quality and granularity of the data it interacts with and the robustness of evaluation metrics employed. To explore the challenges and opportunities presented in the identification of personal information using an LLM, we have undertaken the construction of dedicated datasets and have chosen a suite of evaluation metrics. This chapter delves into the nuances of these datasets and the rationale behind the chosen evaluation techniques. Three datasets were meticulously curated, each with its distinct focus.

TABLE I. Statistical Information of the Datasets

| | Dataset I | Dataset II | Dataset III |
|---|---|---|---|
| **Samples** | 180 | 120 | 300 |
| **Positive Samples** | 90 | 60 | 150 |
| **Negative Samples** | 90 | 60 | 150 |
| **Avg. Tokens** | 445.3 | 517.3 | 474.1 |

Dataset I: This dataset consists of 180 samples that specifically contain either personal names or identification codes. These could be anything from National Identity Card numbers to Driver's License numbers and Passport numbers. Aiming for balance, the dataset is split evenly, with 90 positive samples (those containing the said personal information) and 90 negative samples (those devoid of any such information).

Dataset II: A slightly smaller dataset with 120 samples, it zeros in on data containing phone details, home addresses, and family member information. Again, ensuring a balanced dataset, 60 samples are positive, possessing the desired details, while the other 60 are negative samples, which are bereft of such details.

Dataset III: The most comprehensive of the lot, this dataset contains 300 samples that amalgamate the features of both the aforementioned datasets. It includes personal names, identification codes, telephone details, home addresses, and family member information. For consistency and to avoid introducing bias, this dataset is also balanced, with 150 positive and 150 negative samples.

The veracity and reliability of any system's output hinge on the thoroughness of its evaluation. In the domain of personal information detection, false negatives can lead to data breaches, while false positives can unnecessarily flag benign data, leading to inefficiencies. To navigate this delicate balance, a set of robust evaluation metrics was chosen:

Accuracy: This fundamental metric provides a broad overview of the model's performance. It is the ratio of correctly predicted samples to the total samples. Given the balanced nature of the datasets, accuracy offers a straightforward measure of overall performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision delves deeper into the true positives (data correctly identified as containing personal information). It calculates the ratio of true positives to the sum of true positives and false positives (data incorrectly flagged as containing personal information). High precision is indicative of fewer false alarms.

$$Precision = \frac{TP}{TP + FP}$$

Recall: While precision focuses on the accuracy of positive identifications, recall emphasizes the model's ability to identify all potential positives. It is the ratio of true positives to the sum of true positives and false negatives (data that contains personal information but was missed by the model). A high recall ensures that fewer instances of personal information go undetected.

$$Recall = \frac{TP}{TP + FN}$$

F1 Score: Balancing precision and recall, the F1 score is the harmonic mean of the two. It provides a single metric that encapsulates the trade-offs between precision and recall. In contexts where both these metrics are equally important, the F1 score becomes a crucial evaluation tool.

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

The significance of a well-constructed dataset and a robust evaluation strategy cannot be understated in the realm of personal information identification. By designing three comprehensive datasets and employing a suite of evaluation metrics, we ensure that our methodologies, when applied, undergo rigorous testing. Such rigorous evaluation paves the way for robust, reliable systems capable of safeguarding personal data with utmost efficacy. As the challenge of personal information protection becomes more pronounced in the digital age, these datasets and metrics will prove indispensable in driving forward innovations and refining existing systems.

## V. Empirical Results and Analysis

### A. Experimental Settings

In order to handle digital archives containing PII with utmost confidentiality, this study necessitated a controlled environment and therefore was conducted within a secure local area network. For the purposes of this experiment, we elected to utilize the open-source large-scale model ChatGLM2-6B [26]. The ChatGLM2-6B, a bilingual (Chinese and English) corpus-trained large model, is available openly and requires approximately 8GB of VRAM for deployment and inference, which was within the capacity of the archival institution's hardware infrastructure. The model was implemented using the PyTorch framework with the Huggingface pipeline for deployment. Notably, the large model was utilized as is, without any prior training or fine-tuning for PII detection tasks, thus operating under a completely unsupervised setting.

The reason we did not directly use the GPT-4 API is due to the archival environment's network constraints. Being an internal local area network (LAN), it lacks the capability to utilize external APIs. Moreover, employing the GPT-4 API posed risks of potential PII leakage from the archives. Considering these factors, we opted to deploy an open-source

Large Language Model within the confines of the local network. This approach not only adhered to the network limitations but also mitigated the risks associated with transmitting sensitive archival data over external networks.

To assess the effectiveness of the ChatGLM2-6B model, a baseline model based on BERT (Bidirectional Encoder Representations from Transformers) was designed. This baseline model underwent a fine-tuning process specifically on Dataset III, after which it was evaluated across three different datasets (Dataset I to III). The distribution for training and testing sets was established at an 80-20 split, respectively. The primary objective of the training was to ascertain whether a record contained PII, framing it as a binary classification problem. The learning rate was set at 1e-5, and an early stopping mechanism was employed to mitigate the risk of overfitting.

This experimental setup provided a stringent test bed for both the unsupervised large model and the supervised baseline, facilitating a comprehensive evaluation of their respective capacities to detect PII within digital records under secure and controlled conditions.

### B. *Experimental Results*

Our experimental results are summarized in Table 2, where we present the performance of different PII detection approaches across three datasets. The BERT-finetune model, having been supervised with training data, serves as a benchmark for comparison against the unsupervised LLM in various configurations.

The LLM-Doc approach involved a direct assessment of full-text archival content by the LLM, using a straightforward prompt to determine if personal privacy information is present. In contrast, the LLM-Sent approach deconstructed the archival content into individual sentences, each assessed separately by the LLM under the same prompt. This division into sentences was implemented to test the hypothesis that a more granular analysis could potentially enhance the precision of PII detection, as it allows the LLM to focus on smaller, more manageable units of text. This could be particularly effective in complex documents where PII is sparsely distributed. However, this segmentation also presents challenges, as it may disrupt the contextual flow present in the full document. The loss of broader context could potentially reduce the LLM's ability to accurately identify and classify PII, especially in cases where understanding the overall narrative or thematic structure of the document is crucial. To address these issues, the LLM-Doc-CRSRF and LLM-Sent-CRSRF methods were developed using the CRSRF framework introduced in Section 3.3. These methods enabled the LLM to not only determine the presence of PII within the whole archival content or each sentence, respectively, but also to identify the types of PII present and provide reasons for its judgments.

The empirical results presented in Table 2, reflecting the accuracy (Acc), precision (P), recall (R), and F1 score (F1) of the models, highlight the strengths and limitations of both approaches. They underscore the trade-offs between the depth of contextual analysis afforded by the LLM-Doc method and the focused, granular examination enabled by the LLM-Sent approach. These findings provide valuable insights into the optimal structuring of textual data for PII detection using LLMs in archival research.

TABLE II. EMPIRICAL RESULTS OF PII DETECTION

| Model | Dataset | Acc. | P | R | F1 |
|---|---|---|---|---|---|
| **BERT-finetune (supervised)** | Dataset I | 87.78% | 89.53% | 85.56% | 87.50% |
| | Dataset II | 85.00% | **90.38%** | 78.33% | 83.93% |
| | Dataset III | 88.00% | **91.30%** | 84.00% | 87.50% |
| **LLM-Doc** | Dataset I | 87.22% | 89.41% | 84.44% | 86.86% |
| | Dataset II | 81.67% | 82.76% | 80.00% | 81.36% |
| | Dataset III | 80.33% | 82.27% | 77.33% | 79.73% |
| **LLM-Sent** | Dataset I | 88.33% | 88.76% | 87.78% | 88.27% |
| | Dataset II | 83.33% | 85.71% | 80.00% | 82.76% |
| | Dataset III | 81.33% | 80.92% | 82.00% | 81.46% |
| **LLM-Doc-CRSRF** | Dataset I | 92.78% | 91.40% | **94.44%** | **92.90%** |
| | Dataset II | 87.50% | 86.89% | 88.33% | 87.60% |
| | Dataset III | 88.33% | 89.66% | 86.67% | 88.14% |
| **LLM-Sent-CRSRF** | Dataset I | **92.22%** | **95.24%** | 88.89% | 91.95% |
| | Dataset II | **89.17%** | 89.83% | **88.33%** | **89.08%** |
| | Dataset III | **89.33%** | 89.86% | **88.67%** | **89.26%** |

The BERT-finetune (supervised) model exhibited strong performance across all datasets, demonstrating the effectiveness of supervised learning for PII detection with F1 scores consistently above 83%. The unsupervised LLM-Doc and LLM-Sent methods, without the benefit of task-specific training, showed a notable decrease in performance compared to the BERT-finetune model, particularly in precision and recall across Dataset II and Dataset III.

However, when employing the CRSRF framework with both LLM-Doc-CRSRF and LLM-Sent-CRSRF approaches, we observed a substantial improvement in all metrics. Notably, the F1 score saw a significant increase, with the LLM-Doc-CRSRF and LLM-Sent-CRSRF configurations outperforming the supervised BERT-finetune model in Dataset I. The enhanced performance of the CRSRF-enhanced methods can be attributed to the structured prompting, which provided a clearer context for the LLM to identify and rationalize PII presence. This improvement underscores the potential of structured relevance feedback in enhancing the PII detection capabilities of unsupervised models.

### C. *Case Study: Interpretable and Unsupervised PII Detection with LLM*

In assessing the practicality of PII detection tools for archival workers, interpretability and the ease of integration into current workflows are key considerations. Our comparative case study not only demonstrates the efficiency of detection models but also emphasizes the enhanced interpretability and unsupervised nature of an LLM in identifying PII.

TABLE III. COMPARATIVE ANALYSIS OF PII DETECTION BETWEEN STANDARD CLASSIFIER AND LLM

| Case | Content (with PII redacted) | BERT result | LLM result |
|---|---|---|---|
| Positive Case | *We are pleased to formally introduce Jack Zhang, who will commence employment at the Third Textile Factory. Mr. Zhang hails from Cixi, Zhejiang, and* | TRUE | hasPII: True,<br>name: Jack Zhang,<br>Identification Code: XXXXXXXXXXXXXX,<br>Address: Hangzhou City XXX district |

|  |  |  |  |
|---|---|---|---|
|  | *currently resides in the XXX community on XXX street in XXX district, Hangzhou. He has been an esteemed member of our team at Bobo Textile Factory for six years. With his higher education background, notable work capacity, and consistently positive attitude, Mr. Zhang has demonstrated exemplary performance. It is with great anticipation that we welcome him to his new role.* |  | XXX street XXX community, **reason: This archival content has PII because it contains identifiable personal details such as name, ID number, and address.** |
| Negative Case | *In our general assembly meeting, it was discussed that employee performance metrics have improved over the last quarter. Strategies for maintaining high productivity levels in the coming months were also deliberated. This year's employee innovation program has sparked numerous creative ideas which have the potential to streamline operations.* | FALSE | hasPII: False, **reason: The text does not contain identifiable personal information such as names, identification numbers, addresses, or other personal data that can be used to trace the identity of an individual.** |

*1) Interpretable Results for Archival Workers:* The first example in our case study, the Positive Case, illustrates the LLM's interpretability advantage. Unlike the Standard Classifier, the LLM provides a detailed rationale behind its identification of PII. It pinpoints the exact PII elements such as name, identification code, and address, and importantly, explains the reasoning behind the classification. This granular level of detail offers archival workers a clear understanding of why certain content is flagged, thereby aiding in the nuanced decision-making process required for sensitive information management.

*2) Unsupervised Detection: Ready Out-of-the-Box:* A standout feature of the LLM is their unsupervised nature. The LLM deployed in this study operates without prior training on a PII detection-specific dataset, demonstrating the power of pre-existing language models in understanding and processing complex privacy-related tasks. This is a substantial benefit for archival institutions, as it removes the need for extensive model training and data labeling which can be resource-intensive and impractical, especially in settings sensitive to privacy concerns.

In the Negative Case, the LLM's result once again underscores its robust unsupervised detection capabilities. It accurately discerns the lack of PII and confidently backs its classification with a valid explanation. This is achieved without the model ever having been explicitly trained to identify what constitutes PII, showcasing an intuitive understanding of privacy-related context—a valuable asset in archival settings where new types of sensitive information may frequently emerge.

The case study solidifies the LLM's role as a valuable tool for archival professionals. Its interpretability facilitates a better understanding of content sensitivity, and its unsupervised nature ensures it is a ready-to-use solution without the prerequisite of a curated training dataset. These characteristics potentially reduce the barrier to entry for deploying advanced PII detection tools in archival systems, promoting both efficiency and adherence to privacy standards.

## VI. Conclusion

The present study embarked on an exploration of the capabilities of LLMs in identifying PII within archival content. The research pivoted on empirical methods to substantiate the feasibility of deploying an LLM for PII detection tasks, culminating in a body of evidence that supports their application.

One of the pivotal findings of this investigation is the satisfactory performance level of the LLM in recognizing PII, an achievement marked without the scaffolding of supervised learning techniques. This breakthrough underlines the potential of unsupervised methods, like those intrinsic to LLMs, to parallel and sometimes even surpass the benchmarks set by traditional, supervised models. LLM, thus, emerge as a promising alternative that can alleviate the heavy reliance on extensive labeled datasets and manual feature engineering, which are typical of supervised approaches.

The case studies detailed within the paper serve as testaments to the interpretative prowess of the LLM. It was observed that the LLM are not only adept at detecting PII but are also capable of providing cogent explanations for their outcomes. This interpretability is crucial; it enhances trust in automated systems and provides insights into their decision-making processes, thereby making an LLM more approachable and understandable to archivists.

In conclusion, the application of LLMs for PII detection in archival documents represents a significant stride forward. The unsupervised nature of these models, coupled with their ability to offer explanations for their results, sets the stage for a new era of archival management where efficiency coalesces with reliability, ushering in robust data protection practices that can keep pace with the ever-evolving digital landscape.

## References

[1] L. Jaillant and A. Rees, "Applying AI to digital archives: trust, collaboration and shared professional ethics," Digital Scholarship in the Humanities, vol. 38, no. 2, pp. 571–585, Jun. 2023.

[2] T. Hutchinson, "Natural language processing and machine learning as practical toolsets for archival processing," Records Management Journal, vol. 30, no. 2, pp. 155–174, Jan. 2020.

[3] J. Schneider et al., "Appraising, processing, and providing access to email in contemporary literary archives," Archives and Manuscripts, vol. 47, no. 3, pp. 305–326, Sep. 2019.

[4] L. Jaillant, "How can we make born-digital and digitised archives more accessible? Identifying obstacles and solutions," Arch Sci, vol. 22, no. 3, pp. 417–436, Sep. 2022.

[5] B. B. Borden and J. R. Baron, "Opening up dark digital archives through the use of analytics to identify sensitive content," in 2016 IEEE International Conference on Big Data (Big Data), Dec. 2016, pp. 3224–3229.

[6] R. Marciano, W. Underwood, M. Hanaee, C. Mullane, A. Singh, and Z. Tethong, "Automating the Detection of Personally Identifiable Information (PII) in Japanese-American WWII Incarceration Camp Records," in 2018 IEEE International Conference on Big Data (Big Data), Dec. 2018, pp. 2725–2732.

[7] T. Hutchinson, "Protecting privacy in the archives: Preliminary explorations of topic modeling for born-digital collections," in 2017

IEEE International Conference on Big Data (Big Data), Dec. 2017, pp. 2251–2255.

[8] T. Hutchinson, “Protecting Privacy in the Archives: Supervised Machine Learning and Born-Digital Records,” in 2018 IEEE International Conference on Big Data (Big Data), Dec. 2018, pp. 2696–2701.

[9] C. Brown and C. Morisset, “Simple and Efficient Identification of Personally Identifiable Information on a Public Website,” in 2022 IEEE International Conference on Big Data (Big Data), Dec. 2022, pp. 4246–4255.

[10] M. Zulqarnain, A. K. Z. Alsaedi, R. Ghazali, M. G. Ghouse, W. Sharif, and N. A. Husaini, “A comparative analysis on question classification task based on deep learning approaches,” PeerJ Comput. Sci., vol. 7, p. e570, Aug. 2021.

[11] da Silva, Carlos Jorge,Augusto Pereira, "Detecting and Protecting Personally Identifiable Information through Machine Learning Techniques." Order No. 29112297, Universidade do Porto (Portugal), Portugal, 2020.

[12] P. Kulkarni and C. N. K, “Personally Identifiable Information (PII) Detection in the Unstructured Large Text Corpus using Natural Language Processing and Unsupervised Learning Technique,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 9, Art. no. 9, Sep. 2021.

[13] L. Jaillant and A. Rees, “Applying AI to digital archives: trust, collaboration and shared professional ethics,” Digital Scholarship in the Humanities, vol. 38, no. 2, pp. 571–585, Jun. 2023.

[14] T. Wu et al., “A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development,” IEEE/CAA Journal of Automatica Sinica, vol. 10, no. 5, pp. 1122–1136, May 2023.

[15] T. Brown et al., “Language Models are Few-Shot Learners,” in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020, pp. 1877–1901. Accessed: Nov. 03, 2023. [Online]. Available: https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[16] T. Kojima, S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, “Large Language Models are Zero-Shot Reasoners,” ArXiv, May 2022, Accessed: Nov. 01, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Large-Language-Models-are-Zero-Shot-Reasoners-Kojima-Gu/e7ad08848d5d7c5c47673ffe0da06af443643bda.

[17] T. Wu et al., “A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development,” IEEE/CAA Journal of Automatica Sinica, vol. 10, no. 5, pp. 1122–1136, May 2023.

[18] L. Sweeney, “k-anonymity: a model for protecting privacy,” Int. J. Unc. Fuzz. Knowl. Based Syst., vol. 10, no. 05, pp. 557–570, Oct. 2002.

[19] J. Wei et al., “Chain of Thought Prompting Elicits Reasoning in Large Language Models,” ArXiv, Jan. 2022, Accessed: Nov. 01, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Chain-of-Thought-Prompting-Elicits-Reasoning-in-Wei-Wang/1b6e810ce0afd0dd093f789d2b2742d047e316d5.

[20] L. Jaillant and A. Caputo, “Unlocking digital archives: cross-disciplinary perspectives on AI and born-digital data,” AI & Soc, vol. 37, no. 3, pp. 823–835, Sep. 2022.

[21] J. Li, R. Zhao, Y. He, and L. Gui, “OverPrompt: Enhancing ChatGPT Capabilities through an Efficient In-Context Learning Approach.” arXiv, May 24, 2023.

[22] S. Wang et al., “GPT-NER: Named Entity Recognition via Large Language Models.” arXiv, Oct. 07, 2023.

[23] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing.” arXiv, Jul. 28, 2021.

[24] S. Ramlochan, “The CRISP Prompt Engineering Method: A Dynamic Framework for Advanced AI Reasoning and Decision-Making,” Prompt Engineering. Accessed: Nov. 03, 2023. [Online]. Available: https://promptengineering.org/the-crisp-method-a-dynamic-framework-for-advanced-ai-reasoning-and-decision-making/.

[25] S. Min et al., “Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?” arXiv, Oct. 20, 2022.

[26] Z. Du et al., “GLM: General Language Model Pretraining with Autoregressive Blank Infilling.” arXiv, Mar. 17, 2022.