Sensitive data detection in structured datasets using large language models

Atul Anand
atulanandsh@gmail.com

Abstract. Detecting personal identifiable information (PII), personal healthcare information(PHI) and similar sensitive data within structured datasets is imperative for various reasons. This has become an important problem for organizations to solve automated detection especially amidst high volumes of data collection with the rise of data protection laws and user privacy importance. Multiple sensitive data detection methods exist in structured and unstructured datasets. Structured data pose a greater challenge for such problems due to lack of external context and general understanding of the data. We describe the importance of the problem's existence. We list out existing solutions in the space with their limitations and direct towards the solution using generative artificial intelligence. We want to focus on how the latest advancements in LLMs can provide a truly global solution to the global problem such as sensitive data detection.

Introduction

American Cryptologist Bruce Schneier once said…
"Data is the pollution problem of the information age, and protecting privacy is the environmental challenge."

With the exponential proliferation of data, organizations are grappling with vast amounts and diverse types of data, including personally identifiable information (PII). PII refers to data that can uniquely identify, contact, or locate an individual. The task of identifying and safeguarding sensitive data on a large scale has become increasingly intricate, costly, and time-intensive. Companies must comply with data privacy regulations such as GDPR and CCPA, necessitating the meticulous identification and protection of PII to ensure adherence to these standards. Sensible measures involve identifying sensitive data, such as names, Social Security Numbers (SSNs), addresses, emails, driver's license details, and more. However, even after identification, the process of implementing safeguards like redaction, masking, or encryption across a broad spectrum of data remains cumbersome.

"Data is the new oil" has become an overused cliche. This highlights the recognition of data as a valuable asset sought after by entrepreneurs to yield significant profits. Initially raw, it requires refinement to unlock its value. We now inhabit a vast digital economy where individuals are essentially reduced to data points. Data holds superiority over subjective opinions, valued for its reliability and predictability. Leveraging existing data enables us to forecast outcomes, gain insights for improved business

performance, and develop effective strategies. It does go into the applications and information flow and most recently the focus has become on building insights from these datasets, building statistical models and machine learning models require adequate data as per the problem. One early account is of an ecommerce organization that was able to deduce and suggest pregnancy based on shopping behavior and was able to even predict the due date to a small window [35].
The models built on personal information in essence do not directly hold the information of the user but are capable of leaking such information. This is also a part of ethical and responsible AI which should not leak personal information. This gave rise to concepts of federated machine learning but it comes with its own set of limitations. There has been multiple research showing such adversarial and injection attacks on specifically instructed models(to not reveal personal information) to reveal actual persons details. Even chatgpt and LLMs have been shown to reveal the exact correct information and address of real people [39][40][41]. However, mishandling data can lead to catastrophic consequences. While inherently potent, data necessitates legal oversight for regulation, giving rise to data protection and privacy laws.
To add to the complexity of the PII detection problem, the data governance laws are not all the same, they more or less differ in definitions, properties, region wise, agewise etc.

Complexity of data governance laws: Data privacy laws can differ significantly between countries and regions due to variations in legal frameworks, cultural norms, and governmental priorities. Some laws, like the European Union's General Data Protection Regulation (GDPR), have extraterritorial reach, applying to any organization that processes personal data of EU residents, regardless of where the organization is located. Others may only apply within specific jurisdictions.The California Consumer Privacy Act (CCPA) applies to businesses that meet certain criteria and have customers in California. Another example is DADP act applicability to non residents - "The DPDP Act applies to Indian residents and businesses collecting the data of Indian residents. Interestingly, it also applies to non-citizens living in India whose data processing "in connection with any activity related to offering of goods or services" happens outside India. This has implications for, say, a U.S. citizen residing in India being provided digital goods or services within India by a provider based outside India."[1] The bill introduced an entity known as "consent managers," who were intermediaries for collecting and providing consent to businesses on behalf of individuals. [1] The bill grouped personal data into different categories and required elevated levels of protection for "sensitive" and "critical" personal data. [1] The DPDP Act also states that any processing that is likely to have a detrimental effect on a child is not permitted. The law prohibits tracking, behavioral monitoring, and targeted advertising directed at children.18 The government can prescribe exemptions from these requirements for specific purposes. [1]

Different laws may have varying definitions of what constitutes personal data. For example, some laws may include identifiers like IP addresses and device IDs, while others may not. Some data protection laws, like the GDPR, define personal data broadly to include any information relating to an identified or identifiable individual. In contrast, other laws may have narrower definitions, excluding certain types of data like anonymized information. Data privacy laws may differ in their requirements for obtaining consent from individuals to collect, process, and share their personal data. Some laws may require explicit, opt-in consent, while others may allow for implied consent. For example - The GDPR requires organizations to obtain explicit consent from individuals before processing their personal data for specific purposes. In contrast, the United States does not have a comprehensive federal privacy law, and consent requirements may vary by state or sector. Laws differ in the rights they afford to individuals regarding their personal data. These rights may include the right to access, rectify, delete, and restrict the processing of personal data. Requirements for reporting data breaches to regulatory authorities and affected individuals vary between jurisdictions. Some laws mandate prompt notification, while others may have more lenient reporting timelines. Laws may have different mechanisms and requirements for transferring personal data across borders, including adequacy decisions, standard contractual clauses, binding corporate rules, and certification mechanisms. Enforcement mechanisms and penalties for non-compliance can vary significantly. Some laws empower regulatory authorities to impose substantial fines, while others may rely on alternative forms of enforcement such as audits, warnings, or corrective actions. Some jurisdictions have sector-specific data privacy regulations that apply to particular industries such as healthcare, finance, or telecommunications, in addition to general data protection laws. Example - The Health Insurance Portability and Accountability Act (HIPAA) in the United States imposes specific data privacy and security requirements on healthcare organizations and their business associates, separate from general data protection laws. Cultural attitudes towards privacy and data protection can influence the development and enforcement of data privacy laws. Countries with a strong emphasis on individual privacy rights may have more stringent regulations. [37] The paper discusses the potential misuse or exploitation of privacy laws, such as the GDPR, to facilitate identity theft or other malicious activities. It explores how certain provisions or requirements of privacy laws might inadvertently create opportunities for cybercriminals to exploit loopholes or vulnerabilities in data protection frameworks. The paper delves into scenarios where cybercriminals can manipulate privacy laws to gain access to personal data or exploit individuals' privacy rights for fraudulent purposes. It discusses strategies or techniques that adversaries might employ to abuse privacy regulations for nefarious ends, highlighting potential weaknesses in the legal frameworks designed to protect individuals' privacy.

Context in data: The structured data pose a special challenge to PII values since they have limited context. A table values are a bunch of values put against each which make sense when looked at with schema and schema definition and externally providing context. This is different from text data or images data. An image can have local context of the nearby pixels and objects in the image which can be enlarged upon to global context with external images. With textual datasets the location of the words or values matter in a sentence, a paragraph and in the document and latent definitions. With large corpuses or such datasets they also have explicit meaning and multiple contexts. This is why it’s possible to create large language or image models and fine tune them for local use and is not easy to do for tabular datasets. Many organizations resort to manual and labor-intensive methods to identify and tag PII within their databases, data warehouses, and data lakes. This approach is prone to errors and consumes significant time, leaving sensitive data inadequately protected and susceptible to regulatory fines and security breaches. Sensitive data detection design should be such that it can be used within pipelines of data transfer and not only when it is persisted on a storage server. If the solution is automatic it can be useful as stream level PII monitoring.

Here are some examples of organizations paying huge fines.

Date	Organization	Amount	Issued by	Reason(s)
2018-10	Hospital do Barreiro	€400,000	Portugal (CNPD)	"...based on access policies to databases, which allowed technicians and physicians to consult patients' clinical files, without proper authorization."
2019-03-07	Unnamed bank	€1,560	Hungary (NAIH)	Failure to erase and correct data at the request of the data subject.
2019-04-04	Rousseau (participatory democracy platform)	€50,000	Italy (GPDP)	Failing to protect users' personal data.
2019-03-16	Lower Silesian Football Association	€13,000	Poland (UODO)	Listing personal information of 585 referees on its website.
2019-06	La Liga	€250,000	Spain (AEPD)	Poorly disclosing purpose for requesting GPS and microphone permissions within the football league's mobile app. When the app was open, it transmitted the user's location if it detected an acoustic fingerprint embedded within game telecasts. This was used to help pinpoint the locations of venues that may be screening the games from unauthorized feeds.
2019-06-11	IDDesign A/S (furniture)	DKK 1,500,000	Denmark (Datatilsynet)	Failure to delete personal data from an older system: processing personal data for a longer time than necessary.
2019-06-18	Sergic (real estate services)	€400,000	France (CNIL)	Failure to implement appropriate security measures; failure to define appropriate data retention periods for the personal data of unsuccessful rental candidates.

2019-06-27	UniCredit Bank Romania	€130,000	Romania (ANSPDCP)	Failure to implement appropriate technical and organizational measures.
2020-10-30	Marriott International	£18,400,000	UK (ICO)	Failure to keep millions of customers' personal data secure.
2019-07-08	British Airways	£183,000,000	UK (ICO)	Use of poor security arrangements that resulted in a 2018 web skimming attack affecting 500,000 consumers.Was later reduced to £20 million.
2019-12-09	1&1 Ionos	€9,550,000	Germany (BfDI)	Insufficient protection of personal data, failing to put "sufficient technical and organizational measures" in place to protect customer data in its call centers. Violation of article 32 of GDPR.
2020-12-10	Amazon Europe Core Sarl	€35,000,000	France (CNIL)	Deposit of cookies without obtaining consent and lack of information provided to users.
2020-12-10	Google LLC	€60,000,000		Deposit of cookies without obtaining consent, lack of information provided to users and defective "opposition" mechanism.
2020-12-10	Google Ireland Limited	€40,000,000		
2023-05-12	Meta Platforms	€1.2 billion	Ireland	Transferring data from the European Union to the United States without adequate privacy protections.
2021-12-16	Psykoterapiakeskus Vastaamo	€608,000	Finland	Failure to protect sensitive medical data.

2021-04-15	Vodafone Espana, S.A.U.	€150,000 (reduced to €90,000)	Spain (AEPD)	Violation of Article 6(1)(a) GDPR by processing personal data without consent or any other legal basis. When imposing the fine, the AEPD took into account: ● The type of data affected: basic identifiers such as names, surnames, phone number. ● The relation between the processing and the business activities of the respondent. ● The previous fines on the same grounds. ● The lack of diligence regarding the erasure request. The AEPD finally fined Vodafone €150,000, that was reduced to €90,000 due to the assumption of responsibility and the early payment.
2021-03-24	CP&A B.V.	€15,000	The Netherlands (AP)	Violation of Article 4(15) GDPR, Article 9 GDPR and Article 32 GDPR by processing the health data of sick employees, and for failing to implement appropriate security measures regarding such processing.
2021-03-18	Air Europa Líneas Aéreas S.A.	€600,000	Spain (AEPD)	infringement of Articles 32(1) and 33 GDPR, due to the lack of appropriate technical and organizational measures and of an adequate level of security and due to the delay in the notification of a personal data breach.

Table 1: Companies showing huge fines mainly due to failure of data protection infrastructure or not having one in the first place. [5]

If we look at the reasons they mainly are:

- Unauthorized Access to Personal Data:
  - *Failure to implement proper access controls*, allowing unauthorized personnel to view sensitive information such as patient records.
  - *Lack of authorization protocols* for technicians and physicians accessing clinical files.
- Data Mismanagement and Non-Compliance:
  - *Failure to erase and correct data upon request from data subjects*.
  - Processing personal data for longer durations than necessary without *proper deletion protocols*.
- Inadequate Data Protection Measures:
  - Failing to protect users' personal data on digital platforms.

- *Insufficient implementation of technical and organizational measures* to safeguard personal data.

- Privacy Violations and Poor Security Practices:
  - Poor disclosure of data collection purposes to users.
  - Use of *inadequate security arrangements* resulting in data breaches affecting large numbers of consumers.
  - Failure to notify data breaches in a timely manner as per regulatory requirements.
- Cookie Consent and Information Provision Violations:
  - Deposit cookies without obtaining proper consent from users.
  - Lack of adequate information provided to users regarding data processing activities.
- Data Transfer Violations:
  - Transferring personal data from the European Union to other countries without ensuring adequate privacy protections.
- Health Data Processing Violations:
  - Processing health data without appropriate consent and security measures.
  - Violation of GDPR articles related to *health data processing* and security protocols.

Many reasons are indicative of a missing system or failure to detect and segregate sensitive data. Such a system is necessary for multiple reasons. Firstly, it ensures the protection of individuals' privacy by identifying and securing sensitive information like names, addresses, and financial details. Secondly, compliance with stringent data protection regulations such as GDPR and CCPA necessitates the proper handling of PII within structured datasets to avoid legal penalties. Thirdly, proactive detection of PII mitigates risks associated with data breaches, identity theft, and financial fraud, thereby safeguarding organizations' reputation and financial stability. Additionally, demonstrating a commitment to privacy protection fosters trust among stakeholders, while effective data governance practices ensure transparency, accountability, and integrity in data management processes. In essence, the detection of PII in structured datasets is indispensable for maintaining privacy, complying with regulations, mitigating risks, building trust, and upholding sound data governance principles in today's data-centric environment.

Existing solutions

Many solutions exist for identification and masking of PII from tabular and textual datasets; For textual datasets the solutions mostly use a named entity recognition model on pretrained labels and the labels are fixed. Some solutions in this regard are pii-codex [28], octopii [25], piidetect [26], gretel [11], presidio (from Microsoft) [27], piitools [14].

For tabular solutions also solutions use some kind of classification model, rule based classification, or/and even regular expressions matching for identifying the PII elements.
These kind of models get limited by:

1. Support of countries and region of the dataset as the data governance laws vary
2. The language of the dataset
3. Number of elements/classes they can actually classify
4. Errors in classification - since negative sample space becomes very huge and sparsely categorical

5. Coverage of the data governance laws - which keep on amending and changing with years with new laws in every country

Following solutions are representative of paid and open source solutions in the market for tabular datasets.

a. **Gretel** - The Gretel Classify model is used to identify personal identifiable information (PII) in a tabular data. Classify can detect 40+ Supported Entities including names, addresses, credentials, and more. In this example, we will identify PII from a Sample Dataset containing names, email addresses, phone numbers, credit card numbers, and SSNs. It provides a list of entities it can identify and which regulations those entities are recognize - those are one or more of GDPR, HIPAA, CPRA. [11]
b. **PiiCatcher** - PIICatcher(Tokern) is a scanner for PII and PHI information. It finds PII data in your databases and file systems and tracks critical data. PIICatcher uses two techniques to detect PII:
    i. Match regular expressions with column names
    ii. Match regular expressions and use NLP libraries to match sample data in columns.
c. **PiiTools** - Supports both structured and text data [14]. Also has multi region identifiers [30]. The built-in AI detectors in PII Tools use document context to check whether the detected instance is personal or not. Just because the word "sex" or "white" or "disease" or "Berlin" appear in a document doesn't mean the document is sensitive [31].
d. **Piianalyzer** - uses the following tools for the task:
    i. Common Regular expressions: for extracting some types of 'PII' such as email addresses, phone numbers, street addresses, credit card numbers,
    ii. Stanford Named Entity Tagger: for extracting the locations, organizations and peoples names.[29]
e. **Informatica Claire** - Claire aims at automatic data governance tasks and provisions sensitive data detection as part of data catalog automation[24].
f. **GoogleCloudDlp** - When a dataset hasn't been characterized, Sensitive Data Protection can also inspect the data for sensitive information by using more than 100 built-in classifiers [32][33][34].
g. **AmazonAWS** - Sensitive Data Protection on Amazon Web Services uses Amazon Glue Data Catalog, Amazon Glue Crawler, and the Amazon Glue PII Detection feature as key components to produce data catalogs and perform sensitive data discovery jobs for each Amazon Web Services account. It centralizes information such as data catalogs and data discovery results in a single place and provides a dashboard and detailed reports, making it easier for customers to manage their sensitive data protection efforts. With this information, customers can take appropriate actions to secure sensitive data and comply with regulations such as GDPR, HIPAA, and PIPL. It utilizes machine learning and pattern matching technologies to automatically identify sensitive data. The solution offers built-in sensitive data types for you to choose from. In addition, you can define custom sensitive data types based on business needs. [10]

After identification, redaction, masking, or encryption, cell suppression of sensitive data can be implemented with healthverity, privTAB, etc.

Methods

For the challenges mentioned about the structured dataset the most appropriate method seems to be a character level entity classifier, i.e. character level sequential models (recently recurrent neural networks) appear to be most suited for the solution. Authors in [37] use and compare results on Character-level neural network models including CNN, LSTM, BiLSTM-CRF, and CNN-CRF are investigated on two prediction tasks: (i) entity detection on multiple data formats, and (ii) column-wise entity predic- tion on tabular datasets [37]. Upon testing the implementation we find the following results:

Patients

	Column	Prediction
0	Id	UUID
1	BIRTHDATE	DATE
2	DEATHDATE	DATE
3	SSN	SSN
4	DRIVERS	DRIVERS_LICENSE
5	PASSPORT	DRIVERS_LICENSE
6	PREFIX	UNKNOWN
7	FIRST	UNKNOWN
8	LAST	UNKNOWN
9	RACE	UNKNOWN
10	ETHNICITY	UNKNOWN
11	BIRTHPLACE	UNKNOWN|US_STATE
12	ADDRS	ADDRESS
13	CITY	UNKNOWN
14	STATE	US_STATE
15	COUNTY	UNKNOWN
16	ZIP	FLOAT
17	LAT	FLOAT
18	LON	FLOAT
19	HEALTHCARE_EXPENSES	FLOAT
20	HEALTHCARE_COVERAGE	FLOAT

38% compliance risk

Sales

	Column	Prediction
0	Row ID	INTEGER
1	Order ID	DRIVERS_LICENSE
2	Order Date	DATE
3	Ship Date	DATE
4	Ship Mode	UNKNOWN
5	Customer ID	INTEGER|DRIVERS_LICENSE
6	Customer Name	UNKNOWN
7	Segment	UNKNOWN
8	Country	UNKNOWN
9	City	UNKNOWN
10	State	US_STATE
11	Postal Code	FLOAT
12	Region	UNKNOWN
13	Product ID	DRIVERS_LICENSE
14	Category	UNKNOWN
15	Sub-Category	UNKNOWN
16	Product Name	UNKNOWN
17	Sales	FLOAT

57% compliance risk, 28.5% falsely identified

Internet Attack

	Column	Prediction
0	datetime	DATETIME|FLOAT
1	host	UNKNOWN
2	src	PHONE_NUMBER|BAN
3	proto	UNKNOWN
4	type	ORDINAL
5	spt	FLOAT
6	dpt	FLOAT
7	srcstr	IPV4
8	cc	UNKNOWN
9	country	UNKNOWN
10	locale	UNKNOWN
11	localeabbr	INTEGER
12	postalcode	INTEGER
13	latitude	FLOAT
14	longitude	FLOAT
15	Unnamed: 15	FLOAT

16.67% compliance risk

The models are limited to the dataset they are trained on and on the number of classes they can identify. In an incremental model it will be needed to keep on adding newer labels or progressively adding branches without the use of older dataset in the network [38]. The solutions provided above are using multiple models or classifiers(both machine learning and regex based) to overcome this problem.

The method probably closest to our approach is CASSED (Context based Approach for Structured Sensitive Data Detection), which creates column context by combining column metadata and cell values and integrates them into a single input for a natural language embedding model (BERT). Subsequently, BERT allows the classification of individual columns into one or more labels in a way that the natural language embedding can consider multiple cells of the same column simultaneously. CASSED also employs rule-based methods to aid in the classification of formulaic data types such as social security numbers or credit card numbers.[42][43]

Proof-of-work

When chatgpt was launched it soon became very popular though the gpt models had been developed years before. This was attributed to the reason that for the first time it was released to the people to play with in the form of a chat interface with astounding capabilities. The large language models success comes from multiple factors including getting ability to learn a language from random datasets. When this is trained on huge datasets it can have even more global context than an individual which makes it difficult for someone if it's making up information that is true in the first place. This ability of global context can be applied to learn about the missing context in structured datasets. Let's take names for example - if a column name explicitly says its name but if not then coming from values a russian name may be alien to a program built to identify english names. This is because the joint distribution of English names with characters differs from other regional names. This will work if we get a sample of all such names from the world and train our classifier which some of the solutions do but one can assume a model as big as gpt would have already seen many such names whilst its training along with the context that it was a name in the texts that it appears. So if for example we ask can it suggest some different region names it should be easily able to provide such names.

For example if asked for czech republic male names it outputs:
Jan, Petr, Tomáš, Jiří, Martin, Pavel, Lukáš, Jakub, Michal, Marek

if asked for russian female names:
Anastasia, Ekaterina, Natalia, Olga, Svetlana, Yelena, Marina, Irina, Tatiana, Anna

You
what can "soniya" as word appear to be?

ChatGPT
The word "Soniya" could potentially be a name, a word in a particular language, a brand name, a title, or even a misspelling or variation of another word. Without further context, it's difficult to determine the specific meaning or interpretation of the word "Soniya."

You
which countries this can be a name?

ChatGPT
The name "Soniya" can be found in various countries and cultures around the world. It is a common given name in countries such as India, Pakistan, Bangladesh, and other South Asian countries. Additionally, variations of this name may exist in other regions or cultures.

You
what is the word "soniya"?

ChatGPT
"Soniya" appears to be a given name. It is a common name in various cultures, including India, where it is often spelled as "Sonya" or "Sonia." The exact meaning or origin of the name can vary depending on cultural and linguistic contexts.

ChatGPT
The name "Soniya" could potentially be a name in the following countries:

1. India
2. Pakistan
3. Bangladesh
4. Nepal
5. Sri Lanka
6. Russia (often spelled as "Sonya")
7. Bulgaria
8. Serbia
9. Ukraine

Even though fine tuning can be used for learning sensitive classes on the datasets, we are suggesting that it is not required. Fine tuning will also not be possible because every data will be different in case of structured dataset trying to finetune it to specific labels it would not provide the global advantage on all possible classes. Another reason is we are not expecting users to provide labels of personal information since if they can do that, then the rest of the dataset only follows the rest of the columns, not providing any advantage. The paper "Large Language Models are Zero-Shot Reasoners" sheds light on the under-explored capabilities of LLMs in zero-shot reasoning. It demonstrates that LLMs can perform well on reasoning tasks even without specific training examples, as long as they are prompted effectively. A key takeaway from the paper is that a simple prompt, like "Let's think step by step" before providing answer examples, significantly improves the performance of LLMs on various reasoning benchmarks. This suggests that LLMs might have a high potential for complex reasoning tasks when prompted correctly, even without extensive fine-tuning on specific datasets.[44]. [47][45][46][47][48][49][50][51] advocate and showcase similar zero shot learning capabilities of large language models.

Fetching and understanding of the laws

GPT models already have information on data protection laws and their differences. Following is a snapshot of lengthy output from a gpt model when asked about data protection laws around the world .

```
1. General Data Protection Regulation (GDPR) - European Union
- Applies to all companies processing personal data of EU citizens
- Requires explicit consent for data processing
- Provides individuals with the right to access, correct, and erase their personal data
```

```
2. California Consumer Privacy Act (CCPA) - United States
- Applies to companies that do business in California and collect personal information of California residents
- Requires companies to disclose what personal information they collect and how it is used
- Provides individuals with the right to request deletion of their personal information

3. Personal Information Protection and Electronic Documents Act (PIPEDA) - Canada
- Applies to private sector organizations that collect, use, or disclose personal information in the course of commercial
activities
- Requires organizations to obtain consent for data collection and use
- Provides individuals with the right to access and correct their personal information

4. Data Protection Act 2018 - United Kingdom
- Applies to all organizations that process personal data in the UK
- Requires organizations to obtain consent for data processing
- Provides individuals with the right to access and correct their personal information

5. Personal Data Protection Act (PDPA) - Singapore
- Applies to all organizations that collect, use, or disclose personal data in Singapore
- Requires organizations to obtain consent for data collection and use
- Provides individuals with the right to access and correct their personal information

6. Privacy Act 1988 - Australia
- Applies to all organizations that collect, use, or disclose personal information in Australia
- Requires organizations to obtain consent for data collection and use
- Provides individuals with the right to access and correct their personal information

7. Lei Geral de Proteção de Dados Pessoais (LGPD) - Brazil
- Applies to all organizations that process personal data in Brazil
- Requires organizations to obtain consent for data processing
- Provides individuals with the right to access and correct their personal information

8. Personal Information Protection Act (PIPA) - Japan
- Applies to all organizations that handle personal information in Japan
- Requires organizations to obtain consent for data collection and use
- Provides individuals with the right to access and correct their personal information
```

Offline gpt models for privacy

One might be concerned that sharing data with online LLMs may itself leak privacy. We can take multiple ways to handle it. We sample the very small dataset like 5 to 10 records and shuffle the values and put them through online LLMs. We can do this few times to get general results. Also some providers offer privacy-friendly services in paid plans. Another option is using offline models - dolly[53] and mistral[54] offer leading models in this respect which can be deployed natively to the organization.
Opting for LLMs provides us also in multilingual datasets. We can use multilingual LLMs which can be used to identify not only the languages in the dataset but also the sensitive data in that language. [55]

Sensitivity detection on datasets

Students dataset - UCI ML repository [56] - using gemini version 0.4.1

Columns that look like PII:

- **address:** This column contains the physical address of the individuals, which is considered PII under most data protection laws.
- **famsize:** This column contains the family size of the individuals, which can be used to infer the number of people living in the household.
- **Medu:** This column contains the education level of the mother of the individuals, which can be used to infer the socioeconomic status of the family.
- **Fedu:** This column contains the education level of the father of the individuals, which can be used to infer the socioeconomic status of the family.
- **Mjob:** This column contains the occupation of the mother of the individuals, which can be used to infer the socioeconomic status of the family.
- **Fjob:** This column contains the occupation of the father of the individuals, which can be used to infer the socioeconomic status of the family.

Data protection laws that the given data risks violating:

- **GDPR:** The GDPR defines PII as any information that can be used to identify a natural person, directly or indirectly. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the GDPR.
- **CCPA:** The CCPA defines PII as any information that identifies, relates to, describes, or is capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the CCPA.
- **PIPEDA:** PIPEDA defines PII as any information about an identifiable individual that is recorded in any form. The columns listed above all contain information about identifiable individuals, so the data risks violating PIPEDA.
- **HIPAA:** HIPAA defines PII as any information that is used or disclosed to carry out health care transactions and that relates to the past, present, or future physical or mental health or condition of an individual. The data in the dataset does not contain any health information, so it does not risk violating HIPAA.
- **Data Protection Act 2018:** The Data Protection Act 2018 defines PII as any information that relates to an identified or identifiable living individual. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the Data Protection Act 2018.
- **Privacy Act 1988:** The Privacy Act 1988 defines PII as any information about an individual that is contained in a record that is subject to the Act. The data in the dataset is not contained in a record that is subject to the Privacy Act 1988, so it does not risk violating the Privacy Act 1988.

- **PDPA:** The PDPA defines PII as any data that can be used to identify an individual. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the PDPA.
- **Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011:** The Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011 defines PII as any information that can be used to identify an individual. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011.
- **PIPA:** The PIPA defines PII as any information that can be used to identify an individual. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the PIPA.
- **POPIA:** The POPIA defines PII as any information that can be used to identify an individual. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the POPIA.
- **Data Protection Law:** The Data Protection Law defines PII as any information that can be used to identify an individual. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the Data Protection Law.
- **Personal Data Protection Law:** The Personal Data Protection Law defines PII as any information that can be used to identify an individual. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the Personal Data Protection Law.
- **Law on Protection of Personal Data:** The Law on Protection of Personal Data defines PII as any information that can be used to identify an individual. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the Law on Protection of Personal Data.
- **PDPA:** The PDPA defines PII as any data that can be used to identify an individual. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the PDPA.
- **PECR:** The PECR defines PII as any information that can be used to identify an individual. The columns listed above all contain information that can be used to identify the individuals in the dataset, so the data risks violating the PECR.

The LLM is able to detect PII even without providing the column names. It may be that the model has already seen the dataset while training as it is a public dataset. For this we consider a private dataset which it doesn't have access to. The data was anonymised, randomized and jumbled values for the test. It has following columns and results:

```
['order_id', 'customer_id', 'order_status', 'order_purchase_timestamp',
       'order_approved_at', 'order_delivered_carrier_date',
       'order_delivered_customer_date', 'order_estimated_delivery_date']
```

Columns that look like PII:

- **customer_id:** This column contains unique identifiers for customers, which could be considered PII under GDPR, CCPA, PIPEDA, PDPA, and POPIA.
- **order_purchase_timestamp:** This column contains timestamps of when orders were placed, which could be considered PII under GDPR, CCPA, PIPEDA, PDPA, and POPIA.
- **order_approved_at:** This column contains timestamps of when orders were approved, which could be considered PII under GDPR, CCPA, PIPEDA, PDPA, and POPIA.
- **order_delivered_carrier_date:** This column contains timestamps of when orders were delivered to carriers, which could be considered PII under GDPR, CCPA, PIPEDA, PDPA, and POPIA.
- **order_delivered_customer_date:** This column contains timestamps of when orders were delivered to customers, which could be considered PII under GDPR, CCPA, PIPEDA, PDPA, and POPIA.
- **order_estimated_delivery_date:** This column contains timestamps of when orders were estimated to be delivered, which could be considered PII under GDPR, CCPA, PIPEDA, PDPA, and POPIA.

Laws that the given data risks violating:

- GDPR
- CCPA
- PIPEDA
- PDPA
- POPIA

Laws that the given data does not risk violating:

- HIPAA
- Data Protection Act 2018
- Privacy Act 1988
- PIPA
- Data Protection Law
- Personal Data Protection Law
- Law on Protection of Personal Data
- PDPA
- PECR

Second private dataset has following columns and response:

```
['device_key', 'device_name', 'geo_key', 'date_key', 'order_id',
       'user_key', 'state_name', 'day_name', 'quarter_name', 'month_name',
       'order_time', 'product_key', 'product_name', 'category', 'rating',
       'work_phone', 'gender', 'name', 'email_work', 'email_personal',
       'quantity', 'unit_price', 'age', 'age_group', 'total_valuation',
       'payment_method', 'customer_id', 'campaign_channel', 'visits',
       'device_os', 'traffic_source', 'addressabilty', 'location',
       'geolocation_state', 'order_amount']
```

ColumnName	SensitiveData type	SensitiveBoolean	Explanation	ViolatedLaws
name	Personal Identifiable Information	Yes	Name is a personal identifier.	GDPR, CCPA, HIPAA
email_work	Personal Identifiable Information	Yes	Email address is a personal identifier.	GDPR, CCPA, HIPAA
email_personal	Personal Identifiable Information	Yes	Email address is a personal identifier.	GDPR, CCPA, HIPAA
work_phone	Personal Identifiable Information	Yes	Work phone number is a personal identifier.	GDPR, CCPA, HIPAA
addressabilty	Personal Identifiable Information	Yes	Addressability is a personal identifier.	GDPR, CCPA, HIPAA
location	Personal Identifiable Information	Yes	Location is a personal identifier.	GDPR, CCPA, HIPAA
geolocation_state	Personal Identifiable Information	Yes	Geolocation state is a personal identifier.	GDPR, CCPA, HIPAA

We also tested on synthea careplans dataset which is Synthetic patient and population health data for the state of Massachusetts [57]

ColumnName	SensitiveData type	SensitiveBoolean	Explanation	ViolatedLaws
PATIENT	Personal Identifiable Information (PII)	Yes	This column contains unique identifiers for patients.	GDPR, CCPA, HIPAA, PIPEDA, PDPA, POPIA, PIPA, PECR
ENCOUNTER	Personal Identifiable Information (PII)	Yes	This column contains unique identifiers for encounters.	GDPR, CCPA, HIPAA, PIPEDA, PDPA, POPIA, PIPA, PECR
CODE	Medical Information	Yes	This column contains medical codes that can be used to identify a patient's medical condition.	HIPAA, PIPEDA, PDPA, POPIA, PIPA
DESCRIPTION	Medical Information	Yes	This column contains descriptions of medical procedures that can be used to identify a patient's medical condition.	HIPAA, PIPEDA, PDPA, POPIA, PIPA
REASONCODE	Medical Information	Yes	This column contains reason codes that can be used to identify a patient's medical condition.	HIPAA, PIPEDA, PDPA, POPIA, PIPA
REASONDESCRIPTION	Medical Information	Yes	This column contains descriptions of reasons for medical procedures that can be used to identify a patient's medical condition.	HIPAA, PIPEDA, PDPA, POPIA, PIPA

On synthea patients dataset with columns [57]

```
['Id', 'BIRTHDATE', 'DEATHDATE', 'SSN', 'DRIVERS', 'PASSPORT', 'PREFIX',
       'FIRST', 'LAST', 'SUFFIX', 'MAIDEN', 'MARITAL', 'RACE', 'ETHNICITY',
       'GENDER', 'BIRTHPLACE', 'ADDRESS', 'CITY', 'STATE', 'COUNTY', 'FIPS',
       'ZIP', 'LAT', 'LON', 'HEALTHCARE_EXPENSES', 'HEALTHCARE_COVERAGE',
       'INCOME']
```

ColumnName	SensitiveData type	SensitiveBoolean	Explanation	ViolatedLaws
BIRTHDATE	Date of Birth	Yes	Date of Birth is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
DEATHDATE	Date of Death	Yes	Date of Death is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
SSN	Social Security Number	Yes	Social Security Number is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
DRIVERS	Driver's License Number	Yes	Driver's License Number is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
PASSPORT	Passport Number	Yes	Passport Number is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
PREFIX	Name Prefix	Yes	Name Prefix is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
FIRST	First Name	Yes	First Name is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
LAST	Last Name	Yes	Last Name is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
SUFFIX	Name Suffix	Yes	Name Suffix is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
MAIDEN	Maiden Name	Yes	Maiden Name is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
MARITAL	Marital Status	Yes	Marital Status is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
RACE	Race	Yes	Race is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
ETHNICITY	Ethnicity	Yes	Ethnicity is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
GENDER	Gender	Yes	Gender is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
BIRTHPLACE	Birth Place	Yes	Birth Place is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
ADDRESS	Address	Yes	Address is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
CITY	City	Yes	City is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
STATE	State	Yes	State is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
COUNTY	County	Yes	County is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
FIPS	FIPS Code	Yes	FIPS Code is a sensitive data as it can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (

We also tested on Dessi dataset[42] which has 6272 columns and the snapshot of a result follows:

ColumnName	SensitiveData type	SensitiveBoolean	Explanation	ViolatedLaws
email	Personal Identifier	Yes	Email address can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
ldfncafwhfzv	Personal Identifier	Yes	Government issued ID number.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
email_sid	Personal Identifier	Yes	Email address can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
phonenumber	Personal Identifier	Yes	Phone number can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
pxsgydcjmcmj	Personal Identifier	Yes	Government issued ID number.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
country_name	Personal Identifier	Yes	Country name can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
address	Personal Identifier	Yes	Address can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
nin	Personal Identifier	Yes	Government issued ID number.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
vmodawvvauqe	Personal Identifier	Yes	Government issued ID number.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
mobilephonenumber	Personal Identifier	Yes	Phone number can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
itittmxlzqjr	Personal Identifier	Yes	Government issued ID number.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
dadtysfqjbcs	Personal Identifier	Yes	Government issued ID number.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
zvneublxkuwy	Personal Identifier	Yes	Government issued ID number.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
personal_numerical_code.12	Personal Identifier	Yes	Government issued ID number.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
customer/contact	Personal Identifier	Yes	Name can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
latdeg.12	Personal Identifier	Yes	Latitude can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
genderid.7	Personal Identifier	Yes	Gender can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
ztyhgaddzlun	Personal Identifier	Yes	Government issued ID number.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
geo_latitude.18	Personal Identifier	Yes	Latitude can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR
date_account_closed.6	Personal Identifier	Yes	Date can be used to identify a person.	GDPR, CCPA, PIPEDA, HIPAA, DPA 2018, Privacy Act 1988, PDPA, IT (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011, PIPA, POPIA, Data Protection Law, Personal Data Protection Law, Law on Protection of Personal Data, PDPA, PECR

Conclusion

Detection of personal identifiable information (PII) within structured datasets is a critical challenge for organizations, particularly in the face of evolving data protection laws and increasing privacy concerns. We explored existing solutions and their various methods such as named entity recognition models, classification models, rule-based approaches, and combinations thereof. However, these solutions are often limited by factors such as dataset specificity, language barriers, and evolving data governance laws across different regions and industries. We showed how recent advancements in large language models (LLMs) offer promising opportunities to address these challenges effectively. LLMs have demonstrated remarkable capabilities in zero-shot reasoning, enabling them to understand and reason about complex tasks without extensive fine-tuning on specific datasets. Leveraging LLMs for PII detection in structured datasets can provide a global solution capable of handling diverse data types, languages, and regulatory requirements. We can bypass creating multiple models, regex rules and manually identifying and training classes for the models and solve missing external context to structured datasets.

Moreover, LLMs can facilitate the understanding and interpretation of data protection laws worldwide, offering insights into the nuances and differences among various regulations. In conclusion, harnessing the power of LLMs for PII detection in structured datasets holds great potential for enhancing privacy protection, regulatory compliance, and data governance practices in today's data-centric environment. By leveraging the capabilities of LLMs alongside existing solutions, organizations can build robust and scalable systems for sensitive data detection and safeguarding.

References:

[1] https://carnegieindia.org/2023/10/03/understanding-india-s-new-data-protection-law-pub-90624#:~:text=The%202023%20act%20creates%2C%20for,clearly%20enumerated%20in%20the%20law.
[2] https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf.
[3] http://164.100.47.4/BillsTexts/LSBillTexts/Asintroduced/373_2019_LS_Eng.pdf.
[4] https://blog.ipleaders.in/data-protection-laws-in-india-2/#:~:text=Rights%20of%20the%20individuals%2D%20The,compensated%20for%20any%20data%20breach.
[5] https://en.wikipedia.org/wiki/GDPR_fines_and_notices
[6] https://sprinto.com/blog/gdpr-fines/
[7] https://docs.gretel.ai/gretel-basics/getting-started/examples/discover-pii
[8] https://docs.gretel.ai/create-synthetic-data/models/classify/supported-entities
[9] https://www.amazonaws.cn/en/solutions/technology/security/sensitive-data-protection-on-amazon-web-services/
[10] https://www.amazonaws.cn/en/solutions/technology/security/sensitive-data-protection-on-amazon-web-services/#:~:text=Sensitive%20Data%20Protection%20on%20Amazon%20Web%20Services%20uses%20Amazon%20Glue,each%20Amazon%20Web%20Services%20account.
[11] https://gretel.ai/blog/automate-detecting-sensitive-personally-identifiable-information-pii-with-gretel
[12] https://github.com/tokern/piicatcher
[13] https://documentation.pii-tools.com/
[14] https://pii-tools.com/
[15] https://pii-tools.com/wp-content/uploads/2021/10/PII-Tools-NER-Whitepaper.pdf
[16] https://cloud.google.com/architecture/de-identification-re-identification-pii-using-cloud-dlp
[17] https://aws.amazon.com/blogs/big-data/automatically-detect-personally-identifiable-information-in-amazon-redshift-using-aws-glue/
[18] https://www.sciencedirect.com/science/article/pii/S0957417423004256
[19] https://dl.acm.org/doi/abs/10.1145/3510548.3519377
[20] https://onlinelibrary.wiley.com/doi/abs/10.1002/net.3230220407
[21] https://www.sciencedirect.com/science/article/abs/pii/S037722171100316X
[22] https://arxiv.org/pdf/2012.09597.pdf
[23] https://docs.gretel.ai/create-synthetic-data/models/classify/supported-entities
[24] https://www.informatica.com/blogs/machine-learning-data-catalog.html
[25] https://github.com/redhuntlabs/Octopii
[26] https://github.com/edwardcooper/piidetect
[27] https://microsoft.github.io/presidio/
[28] https://github.com/EdyVision/pii-codex
[29] https://gitlab.math.ubc.ca/tomyerex/piianalyzer
[30] https://documentation.pii-tools.com/#available-pii-types
[31] https://documentation.pii-tools.com/
[32] https://cloud.google.com/architecture/de-identification-re-identification-pii-using-cloud-dlp
[33] https://cloud.google.com/sensitive-data-protection/docs/infotypes-reference
[34] https://cloud.google.com/sensitive-data-protection/docs/high-sensitivity-infotypes-reference
[35] https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
[36] Pavur, J., Knerr, C. and Dionach, L.T.D., GDPArrrrr: Using Privacy Laws to Steal Identities.
[37] Truong, A., Walters, A. and Goodsitt, J., 2020. Sensitive data detection with high-throughput neural network models for financial institutions. arXiv preprint arXiv:2012.09597.
[38] Adhikari, S.S., Singh, S., Rajagopal, A. and Rajan, A., 2019. Progressive fashion attribute extraction. arXiv preprint arXiv:1907.00157.

[39] https://www.engadget.com/a-silly-attack-made-chatgpt-reveal-real-phone-numbers-and-email-addresses-200546649.html
[40] https://www.zdnet.com/article/chatgpt-can-leak-source-data-violate-privacy-says-googles-deepmind/
[41] Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A.F., Ippolito, D., Choquette-Choo, C.A., Wallace, E., Tramèr, F. and Lee, K., Scalable Extraction of Training Data from (Production) Language Models.
[42] Kužina, V., Petric, A.M., Barišić, M. and Jović, A., 2023. CASSED: Context-based Approach for Structured Sensitive Data Detection. Expert Systems With Applications, 223, p.119924.
[43] Kužina, V., Vušak, E. and Jović, A., 2021, September. Methods for Automatic Sensitive Data Detection in Large Datasets: a Review. In *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)* (pp. 187-192). IEEE.
[44] Kojima, T., Gu, S.S., Reid, M., Matsuo, Y. and Iwasawa, Y., 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35, pp.22199-22213.
[45] Wang, Z., Yu, A.W., Firat, O. and Cao, Y., 2021. Towards zero-label language learning. arXiv preprint arXiv:2109.09193.
[46] https://timx.me/blog/2023/computers-vs-llms/
[47] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. Advances in neural information processing systems, 33, pp.1877-1901.
[48] Sanh, V., Webson, A., Raffel, C., Bach, S.H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T.L., Raja, A. and Dey, M., 2021. Multitasking prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207.
[49] Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T. and Kong, L., 2022. Zerogen: Efficient zero-shot learning via dataset generation. arXiv preprint arXiv:2202.07922.
[50] Ateia, S. and Kruschwitz, U., 2023. Is ChatGPT a Biomedical Expert?--Exploring the Zero-Shot Performance of Current GPT Models in Biomedical Tasks. arXiv preprint arXiv:2306.16108.
[51] Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M. and Jiang, Y., 2023. Zero-shot information extraction via chatting with chatgpt. arXiv preprint arXiv:2302.10205.
[52] https://www.kaggle.com/datasets/sensitivedetection/dessi-dataset-for-structured-sensitive-information
[53] https://huggingface.co/databricks/dolly-v2-12b
[54] https://mistral.ai/news/mistral-large/
[55] https://ai.meta.com/tools/fairseq/
[56] Cortez,Paulo. (2014). Student Performance. UCI Machine Learning Repository. https://doi.org/10.24432/C5TG7T
[57] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, Scott McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, Journal of the American Medical Informatics Association, Volume 25, Issue 3, March 2018, Pages 230–238, https://doi.org/10.1093/jamia/ocx079