UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Confidentiality

15–17 October 2025, Barcelona

Detecting Contextually Sensitive Data with AI

Liang Telkamp[1], Melanie Rabier[2], Javier Teran[2], Madelon Hulsebos[1]
[1]Centrum Wiskunde Informatica (CWI) [2]OCHA's Centre for Humanitarian Data

corresponding authors: teran1@un.org and madelon@cwi.nl

Abstract

The rise of data sharing through private and public data portals necessitates more attention to detecting and protecting sensitive data before datasets get published. While research and practice have converged on the importance of documenting Personal Identifiable Information (PII), automatic, accurate and scalable methods for detecting such data in (tabular) datasets are behind. Moreover, we argue that sensitive data detection is more than PII type detection, and methods should consider the more fine-grained context of the dataset and how its publication can be misused beyond the identification of individuals. To guide research in this direction, we present a novel framework for contextual sensitive data detection based on *type contextualization* and *domain contextualization*. For type contextualization, we introduce the *detect-then-reflect* mechanism, in which large language models (LLMs) first detect potential sensitive column types in tables (e.g. PII types such as `email address`), and then assess their actual sensitivity based on the full table context. For domain contextualization, we propose the *retrieve-then-detect* mechanism that contextualizes LLMs in external domain knowledge, such as data governance instruction documents, to identify sensitive data *beyond PII*. Experiments on synthetic and humanitarian datasets show that: 1) the detect-then-reflect mechanism significantly reduces the number of false positives for type-based sensitive data detection, whereas 2) the retrieve-then-detect mechanism is an effective stepping stone for domain-specific sensitive data detection, and retrieval-augmented LLM explanations already provide a useful input for manual data auditing processes more efficient.

1 Introduction

Data sharing underpins reuse across sciences, enterprises, and governments through public and private portals (Borgman and Groth, 2025; Worth et al., 2024; Zhang et al., 2024; Wilkinson et al., 2016; Brickley et al., 2019). This raises the need to detect and protect sensitive data. Such data includes direct identifiers (e.g., names, emails (European Parliament and Council of the European Union, 2016)) as well as indirect or contextual information that becomes risky under certain conditions (OCHA, 2025). For example, facility locations, ethnic identifiers, or military logs may be sensitive depending on geopolitics, populations, or timing. This illustrates the complexity of sensitive data detection (SDD) and the need for nuanced, context-dependent methods.

Beyond operational and ethical concerns, Large Language Models (LLMs) increase the risks of leakage: trained on web data, they may memorize and reproduce sensitive information (Subramani et al., 2023; Worth et al., 2024; Carlini et al., 2021; Lukas et al., 2023). Studies show sensitive data is often underspecified in documentation on sharing portals like HuggingFace (Yang et al., 2024; Akhtar et al., 2024; Worth et al., 2024), motivating improved methods and tools for (semi-)automated protection.

Most work on SDD targets Personally Identifiable Information (PII) (Kužina et al., 2023; Subramani et al., 2023; Akhtar et al., 2024). For unstructured data, PII detection relies on Named Entity Recognition or pattern-matching (Subramani et al., 2023); similar approaches exist for tabular data (Raman, 2001). Yet *sensitivity* extends far beyond explicit PII. Kober et al. (2023) propose a taxonomy of always-, combination-, context-, and value-sensitive data. For instance, `email address` may be harmless in one dataset but identifying in another, causing naïve type-based detection yield false positives. Likewise, hospital `geo-coordinates` may be benign in one region but sensitive in conflict zones (OCHA, 2025), calling for contextualization beyond PII detection.

LLMs offer new potential for sensitivity detection by reasoning over structured data. Studies show they can interpret tables with carefully designed prompts (Fang et al., 2024), achieving strong results in tasks such as text-to-SQL and table retrieval (Fang et al., 2024; Ji et al., 2025). These capabilities suggest that LLMs can leverage schema, values, and external knowledge to interpret tabular dataset semantics (Hulsebos et al., 2019a).

We revisit sensitive data in tabular datasets and introduce *type contextualization* and *domain contextualization* as key to contextual SDD. To put this into practice, we propose two mechanisms. First, *detect-then-reflect* addresses type contextualization: LLMs detect potential PII-type columns, then reflect in context, reducing false positives. Second, *retrieve-then-detect* addresses domain contextualization: LLMs use retrieval-augmented guidelines to identify non-personal sensitive data in expert-informed synthetic datasets, producing effective human-audit explanations. Finally, we evaluate both mechanisms on real humanitarian data from the UN Humanitarian Data Exchange (HDX) platform (United Nations, 2014), showing significant false positive reduction compared to baselines like pattern-matching and basic ML (Microsoft, 2018; Google, 2018).

2 Related work

Conceptualization of Sensitive Data. The notion of "sensitive data" is multifaceted, spanning legal, theoretical, and domain-specific views. Legal frameworks often equate sensitivity with PII. The GDPR defines personal data broadly as "any information relating to an identified or identifiable natural person," specifying sensitive categories like health, ethnicity, or politics requiring extra protection (European Parliament and Council of the European Union, 2016). Similarly, HIPAA defines identifiers that make health information "Protected Health Information (PHI)" when linked to individuals (Centers for Medicare & Medicaid Services, 1996). These provide a baseline: data that directly or indirectly singles out a person is sensitive.

Yet static, category-based definitions are insufficient for modern governance (Quinn and Malgieri, 2021), as they emphasize direct identifiers while overlooking non-personal but risky data such as geolocation data, community-level attributes, or institutional records. Sensitivity is often situational: seemingly harmless data can become sensitive when combined, repurposed, or shared across contexts (Malkin, 2023).

Theoretical perspectives advance this nuance. Nissenbaum's *Contextual Integrity* frames sensitivity as the appropriateness of information flows within social norms (Nissenbaum, 2004). Violations occur when flows breach expectations, not from the data type itself. Building on this, researchers advocate context-grounded assessments (Malkin, 2022; Shvartzshnaider et al., 2019). For instance, shopping data may be ordinary in retail but sensitive in behavioral profiling.
Recent taxonomies further refine this view. Kober et al. (2023) distinguish: (1) *always-sensitive* data (e.g., direct identifiers); (2) *combination-sensitive* (sensitive when linked, e.g., identity + location); (3) *context-sensitive* (sensitivity by use, e.g., photos in medical vs. travel contexts); and (4) *value-sensitive* (sensitivity from specific values). This shows that not all PII is equally sensitive, and not all sensitive data is PII.
In summary, conceptualizations have shifted from static identifiers to layered, context-dependent models. Legal standards remain essential for compliance, but frameworks like Contextual Integrity and multi-layer taxonomies emphasize context, use, and value. This motivates our research: effective sensitive data detection must go beyond PII recognition toward context-aware, domain-sensitive approaches.

Methods for Detecting Sensitive Data. Most automated sensitive data detection methods concentrate on PII recognition. Widely adopted industry tools like Microsoft Presidio (Microsoft, 2018) and Google Cloud DLP (Google Cloud, 2018) are largely rule-based, relying on regular expressions, keyword lists, and simple machine learning heuristics to catch well-defined patterns like email addresses or phone numbers.
Research has advanced towards more semantic understanding. Hulsebos et al. (2019b) introduced *Sherlock*, a deep learning model that classifies table columns into 78 semantic types, including numerous PII categories, by learning from diverse feature sets extracted from column values. This approach outperformed regex and dictionary baselines, especially on noisy or inconsistently formatted data. Kužina et al. (2023) developed this further with *CASSED*, a transformer-based model (BERT) that integrates column headers, data types, and sample values to classify columns into sensitive categories. By incorporating table context, CASSED showed improved generalization and an ability to identify sensitive columns that lack explicit clues in their values.
Recently, Large Language Models (LLMs) have emerged as powerful tools for sensitive data detection (Shen et al., 2025; Yang et al., 2023). Their vast internalized knowledge and semantic understanding allow them to recognize PII even when rigid patterns fail, often in a zero-shot or few-shot manner without task-specific training (Wang et al., 2023; Brown et al., 2020). Studies confirm that LLM-based detectors can outperform both rule-based and traditional machine learning methods in accuracy and adaptability (Shen et al., 2025). However, a key limitation remains: while LLMs excel at identifying known PII types, they struggle with domain-specific *contextual* sensitivity unless explicitly provided with the relevant operational context (Cheng et al., 2024).
This review highlights a significant gap: most research and practical methods focus on PII-type recognition, overlooking non-personal data that becomes sensitive due to its operational, humanitarian, or organizational context. Our work aims to bridge this gap by developing a framework that addresses both personal and non-personal sensitive data detection through contextualization using LLMs.

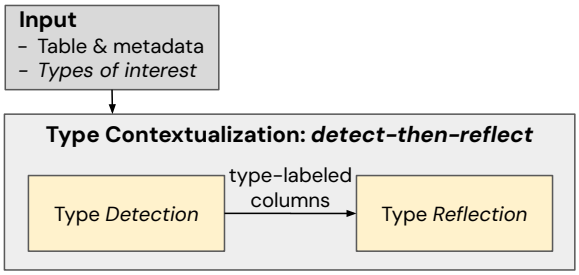
3 Revisiting the Concept of Sensitive Data

The definition of Personally Identifiable Information (PII) is straightforward: it refers to data that can identify an individual (European Parliament and Council of the European Union, 2016). Existing detection methods often recognize standard entities (e.g., names, emails, phone numbers) but treat all occurrences as equally sensitive, leading to high false positive rates. Sensitivity, however, depends on context: an `email address` may reflect personal or organizational use. We call this *type contextualization*, evaluating whether a column truly contains PII requires considering the full table context.
While regulations like GDPR (European Parliament and Council of the European Union, 2016) and HIPAA (Centers for Medicare & Medicaid Services, 1996) emphasize identity-related data, humanitarian contexts demand a broader view of sensitivity. Following Nissenbaum's theory of Contextual Integrity (Nissenbaum, 2004), privacy risks arise when data is used outside its intended setting. Sensitivity may thus extend beyond identity

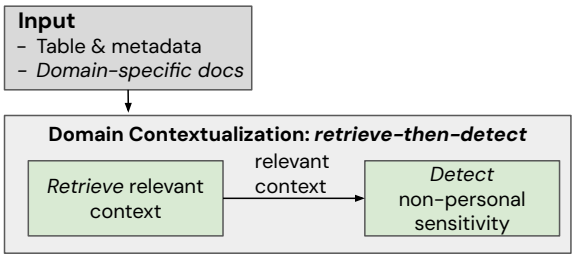
exposure to risks such as targeting or undermining humanitarian access. Similar concerns arise in enterprises or governments, where non-PII (e.g., company-sensitive or classified data) requires protection. We refer to this broader perspective as *domain contextualization*: sensitivity emerges from the dataset's role and potential misuse, not just from the presence of PII.

4 Toward *Contextual* Sensitive Data Detection

Here, we first introduce two mechanisms for contextual sensitive data detection, for type contextualization and domain contextualization. We then provide details of our implementation of these mechanisms.



(A) The *detect-then-reflect* mechanism for detecting sensitive table columns, e.g. PII, using type contextualization to increase the precision. This mechanism entails: ① detecting potentially sensitive columns based on their types, such as `address` and `name`, ② having an LLM reflect on the actual sensitivity of the column in the full table context to, for example, not mark the `address` and `name` of organizations as PII.



(B) The *retrieve-then-detect* mechanism for detecting sensitive data that requires domain contextualization. This mechanism entails: ① retrieving relevant contextual information such as rule-based guidelines, ② having an LLM detect sensitive data using its parametric knowledge, the retrieved instructions, and expert-informed instructions provided in the prompt.

FIGURE 1. Mechanisms for detecting sensitive data.

4.1 Type contextualization: *detect-then-reflect*

Building on our observation that sensitive data detection is a context-specific task, we introduce a two-track pipeline for personal sensitive data: *detect-then-reflect*. This method first performs a broad detection pass to identify all columns that may contain personal identifiers, followed by a context-aware refinement step.
In the initial *detection* phase, potentially sensitive columns are identified using type classification based on column names and representative values. This phase prioritizes recall in order to avoid missing any columns that may carry personal identifiers. In the second *reflection* phase, the model re-evaluates these columns in the context of the full table. We prompt an LLM with a markdown-formatted table, including the column's detected PII type. The model then determines whether the column is contextually sensitive, taking into account inter-column relationships across the entire table. This enables a more precise classification than standard pattern-based methods.

4.1.1 Implementation Detection. We frame PII detection as a multiclass classification task over a taxonomy of 27 entity types (e.g., `PERSON_NAME`, `EMAIL_ADDRESS`, `GENERIC_IDENTIFIER`) derived from GDPR and UN OCHA guidelines (see Appendix A for the complete list). The LLM is prompted with a column name and five sample values and instructed to return only the most appropriate entity label or `None`. This step is designed to maximize high recall.
We employ both zero-shot and fine-tuned approaches. The zero-shot method leverages the model's inherent ability to follow instructions without task-specific training, while the fine-tuned approach enhances performance

by training the Gemma 2 9B and Qwen3 8B base models on a synthetic dataset of 1,000 labeled columns. Fine-tuning is carried out using Low-Rank Adaptation (LoRA) for parameter efficiency, with a rank of $r = 16$ and a scaling factor of $\alpha = 16$ applied to both attention and feedforward layers. The models were optimized with AdamW for a single epoch using an 80/20 train-validation split.

4.1.2 Implementation Reflection. In this phase, the model re-evaluates each candidate PII column within the context of the entire table by being prompted with a markdown-formatted rendering of the full table, including headers and five sample rows per column, along with the specific PII type detected for the target column. It is then instructed to assign one of three contextual sensitivity levels: `NON_SENSITIVE`, referring to information that cannot identify a person (e.g., aggregate data, organization names); `MODERATE_SENSITIVE`, referring to data that could potentially identify a person when combined with other attributes (e.g., demographics, partial information); and `HIGH_SENSITIVE`, referring to data that definitively identifies a person (e.g., full name, email, national ID). This step is crucial for mitigating false positives from the detection stage by incorporating inter-column relationships and table semantics.

4.2 Domain contextualization: *retrieve-then-detect*

While personal data sensitivity can often be inferred from structural context, non-personal sensitive data typically depends on domain-specific and geopolitical factors that lie beyond the data table itself. For such cases, we propose a *retrieve-then-detect* approach.
This method integrates external contextual information, such as country-specific Information Sharing Protocols (ISPs) and policy documents, into the classification process. Relevant documents are retrieved based on the location of a dataset and included in the prompt alongside the table and column under analysis. The LLM leverages both its internal parametric knowledge and the retrieved external document to assess whether a column contains data that should be considered sensitive in the given operational context. Unlike PII detection, where there is a predefined list of entities (e.g. `email`, `phone number`), non-personal sensitive data has no fixed ontology. We retrieve domain-specific column sensitivity based on the broader table context and retrieved contextual knowledge.

4.2.1 Implementation Retrieval. For detecting non-PII sensitive data, we propose a retrieval-augmented generation (RAG) approach that integrates external domain knowledge.
The system retrieves relevant policy fragments based on the dataset's geographical context (e.g., country). For our humanitarian use case, we index country-specific Information Sharing Protocols (ISPs). To reduce prompt length, we use GPT-4o-mini to extract the specific sensitivity rules from these documents (e.g., data types for `LOW`, `MODERATE`, `HIGH`, `SEVERE` sensitivity), which we store in a dictionary. If no country-specific ISP is available, a default fallback template based on cross-country consensus is used.

4.2.2 Implementation Detection. In the detection phase, the LLM is provided with a curated prompt containing table context (headers and five representative rows), relevant ISP extracts containing explicit instructions to classify each column's sensitivity according to the ISP guidelines while justifying its decision by directly citing the relevant policy passages. Based on the policy definitions, the model assigns one of four sensitivity levels, `NON_SENSITIVE`, `MODERATE_SENSITIVE`, `HIGH_SENSITIVE`, or `SEVERE_SENSITIVE`, ensuring that sensitivity assessments are grounded in explicit, auditable domain policies, thereby enhancing both interpretability and practical usability for domain experts.

5 Experimental Setup

This section outlines the datasets, evaluation metrics, baselines, and implementation details used to evaluate our proposed framework for contextually sensitive data detection.

5.1 Datasets

We evaluate our mechanisms on both synthetic and real-world datasets containing PII and non-personal sensitive data, focusing on humanitarian contexts:

- **Synthetic PII data**: 1,000 labeled columns generated with GPT-4o, spanning 27 PII types plus a `None` class. Each column included a name and five values (mix of obvious PII, ambiguous, and misleading non-PII). All examples were manually labeled and used for fine-tuning and zero-shot evaluation.
- **Real PII tables (GitTables)**: 66 anonymized tables (2,061 columns) from GitTables (Hulsebos et al., 2023), annotated for both PII entities and contextual sensitivity to provide ground truth for detect-then-reflect.
- **Synthetic humanitarian tables**: 9 tables based on humanitarian use cases (e.g., displacement tracking, public health) and ISPs. Each contained both safe and sensitive columns (e.g., "access restrictions", "vulnerability score"), with domain experts defining sensitivity labels.
- **Real humanitarian data (placebo set)**: 14 non-sensitive HDX datasets, selected as published, non-sensitive tables from countries with available ISPs. Used to estimate false positive rates.

For domain contextualization (retrieve-then-detect), we evaluated on the 9 synthetic humanitarian tables combined with the 14 non-sensitive HDX datasets.

5.2 Metrics

We evaluate along three axes: PII detection, contextual reflection for PII, and non-PII sensitivity detection. To account for class imbalance, we report both weighted (support-weighted) and macro averages. Weighted averages reflect overall performance across examples, while macro averages treat each class equally and reduce dominance by frequent classes. Unless otherwise noted, we focus on weighted averages, but report macro scores in Appendix, Table 10).

Performance is assessed at both the column and table level. At the column level (type contextualization), PII detection is treated as a multiclass task over 27 PII types (plus a "None" category), while sensitivity reflection reduces to a binary decision between sensitive (moderate/high) and non-sensitive (low). At the table level (domain contextualization), a table is classified as sensitive if it contains any moderate/high/severe sensitive information, resulting in a binary task (sensitive vs. not). Precision, recall, and F1 are used as evaluation metrics. For labeling, sensitivity levels are binarized: *low* is treated as non-sensitive, while *moderate*, *high*, and *severe* are grouped as sensitive. This aligns with operational practice where any moderate or higher sensitivity triggers safeguards, consistent with UN OCHA guidance that such data warrants protective measures (OCHA, 2025).

Finally, alongside classification metrics, we report average inference time per table to assess deployment practicality, particularly where responsiveness and scalability are critical.

5.3 Baselines

We compare our methods against two industry-standard baselines for sensitive data detection. The first is *Google Cloud DLP*, a rule-based system that combines built-in *infoTypes* (predefined detectors for common PII entities) with custom-defined *infoTypes* such as dictionaries and regexes. For our experiments, we configured the minimum likelihood threshold to "Likely". The second baseline is *Microsoft Presidio*, an open-source SDK that leverages regular expressions, pattern-based rules, context windows, and optional NER to detect PII. To enable consistent evaluation, each Presidio entity was mapped to a corresponding category in our unified taxonomy (Appendix, Table 8).

Both systems output column-level predictions and are widely adopted in enterprise settings. However, neither tool supports contextual reflection or table-wide reasoning; any match to a known PII pattern is automatically treated as sensitive.

5.4 Model Selection

We evaluate both proprietary and open-source LLMs. The proprietary GPT-4o-mini (OpenAI et al., 2024) serves as an upper baseline. Open-source models include the Gemma family (2 and 3) (Team, 2025; Team et al., 2024) and Qwen3 in 8B and 14B sizes (Yang et al., 2025). We fine-tuned Qwen3 8B and Gemma 2 9B on our synthetic PII data to create Qwen3 8B-FT and Gemma 2 9B FT.
Qwen3 14B was selected based on its predecessor Qwen2.5's exceptional performance on advanced benchmarks such as MMLU Pro [1], demonstrating strong reasoning capabilities and robust general understanding, particularly among relatively small language models. While Qwen2.5 already showed good performance in its size class, we opted to use the newer Qwen3 variant, anticipating further improvements in reasoning and contextual capabilities. Gemma 2 9B and Gemma 3 12B were included to strike a balance between computational efficiency and performance. These models provide transparency due to their open-source nature and support quantized deployment, making them suitable for scenarios with limited hardware resources or privacy concerns. Aya Expanse 8B serves as a lightweight alternative appropriate for resource-constrained environments, though it benchmarks comparatively lower performance on the MMLU Pro benchmark. Lastly, GPT-4o-mini was incorporated due to its practical performance across various tasks, despite its API-only accessibility, which restricts control over data privacy and deployment flexibility in sensitive use-cases.

Model	**GPU**	**API**	**Params** (B)	**Size** (GB)	**4bit** (GB)	**Supported Tasks**		
						PII Detect	**PII Reflect**	**non-PII Detect**
GPT-4o-mini	✗	✓	-	-	-	✓	✓	✓
GPT-4o	✗	✓	-	-	-	✓	✗	✗
Qwen3 8B	✓	✗	8.0	~16.4	~7.5	✓	✓	✓
Qwen3 14B	✓	✗	14.0	~29.5	~11.1	✓	✓	✓
Gemma 2	✓	✗	9.0	~18.5	~6.1	✓	✓	✓
Gemma 3	✓	✗	12.0	~19.4	~12.8	✓	✗	✗
Aya Expanse	✓	✗	8.0	~16.1	N/A	✓	✓	✓

Table 1. Overview of model characteristics and supported tasks. API-based models (e.g., GPT-4o) are accessed via cloud endpoints, while others require local GPU deployment. Model sizes refer to 16-bit versions; 4-bit quantized sizes are shown where available. Exact parameter size of GPT-4o-mini and GPT-4o are undisclosed by OpenAI. Access is provided only through the API.

6 Results of contextual sensitive data detection

We present a comprehensive evaluation of our proposed framework for contextually sensitive data detection in tabular datasets, focusing on two core mechanisms: Type Contextualization (Detect-Then-Reflect) for PII and Domain Contextualization (Retrieve-Then-Detect) for non-PII sensitivity.

6.1 Type Contextualization: *Detect-Then-Reflect*

6.1.1 LLM-Based PII detection. We evaluated baseline systems (Google DLP, Microsoft Presidio) and a range of LLMs for detecting PII in tabular data. Baselines performed poorly: Google DLP achieved moderately high precision (0.71) but suffered from very low recall (0.32), while Presidio performed worse across all metrics due to schema mismatches and limited use of contextual signals. In contrast, LLMs substantially outperformed

[1] https://huggingface.co/spaces/TIGER-Lab/MMLU-Pro

both baselines. Among zero-shot models, Qwen3 14B and Gemma 3 12B achieved the strongest results, with F1 scores above 0.80. Fine-tuning further improved performance, most notably for Qwen3 8B, which reached an F1 score of 0.94 and outperformed larger models, underscoring the effectiveness of lightweight task-specific adaptation. However, much of this improvement was driven by better handling of the most frequent entity class, `GENERIC_IDENTIFIER`, highlighting challenges in detecting rare PII types. Finally, we tested a privacy-preserving "no-records" configuration that excluded sample values from the prompt. This substantially degraded performance (F1: 0.28 vs. 0.42), indicating that LLMs rely heavily on example values, alongside column names, for accurate classification.

System / Model	Precision	Recall	F1
Google DLP	0.713	0.324	0.298
Presidio	0.198	0.245	0.208
GPT-4o-mini	0.578	0.366	0.422
GPT-4o-mini no records	0.393	0.253	0.284
GPT-4o	0.962	0.658	0.728
Gemma 2 9B	0.960	0.538	0.605
Gemma 2 9B FT	0.945	0.775	0.832
Gemma 3 12B	0.859	0.811	0.818
Qwen3 8B	0.919	0.549	0.582
Qwen3 8B FT	**0.968**	**0.934**	**0.942**
Qwen3 14B	0.931	0.819	0.857
Aya Expanse 8B	0.931	0.394	0.448

(a) Performance of baseline systems (Google Cloud DLP and Microsoft Presidio) and LLMs on PII entity detection in tabular data. Each LLM received a column name and five unique sample values and predicted one of 27 PII types or "None". "FT" denotes a fine-tuned model variant.

Model	Precision	Recall	F1
Ground-truth PII	0.527	**1.000**	0.690
GPT-4o-mini	**0.932**	0.944	**0.938**
Gemma 2 9B	0.705	0.913	0.796
Gemma 3 12B	0.883	0.816	0.848
Qwen3 8B	0.544	0.997	0.703
Qwen3 14B	0.683	0.951	0.795
Aya Expanse 8B	0.552	0.997	0.710

(b) Evaluation of PII sensitivity reflection by LLMs with ground-truth PII entities provided. The first row shows the low precision if all PII columns are labeled sensitive; reflection improves precision significantly with minimal recall loss.

Table 2. Results of contextual sensitive data detection: (a) PII detection and (b) PII sensitivity reflection.

Error analysis revealed that false positives were most often due to date-like values incorrectly labeled as `DATE_OF_BIRTH`, despite non-indicative column names. False negatives primarily arose from the ambiguous `GENERIC_IDENTIFIER` class and from PII encoded in categorical or binary form (e.g., race codes), which masked sensitive meaning. Additional challenges included high-variability entities such as `PERSON_NAME`, `ORGANIZATION_NAME`, `ZIPCODE`, and `STREET_ADDRESS`, where non-standard formats or unfamiliar names reduced recall. Overall, LLMs consistently outperformed rule-based baselines by leveraging semantic cues in both column names and values, but recall remained a key limitation for rare or implicit PII types.

6.1.2 Contextual Sensitivity Reflection. While LLMs detected PII entities effectively, traditional approaches assume all PII is equally sensitive, overlooking contextual factors such as table purpose and structure. To isolate the reflection step, we provided models with ground-truth PII columns and tasked them only with classifying contextual sensitivity based on the full table. This setup removed detection errors and directly evaluated LLMs' ability to reason about sensitivity in context.

Treating all PII-type columns as sensitive achieved perfect recall but low precision, as many fields (e.g., `ORGANIZATION_NAME`, `GENERIC_IDENTIFIER`) were not contextually sensitive. LLMs improved precision dramatically (from 0.53 to >0.90 in the best cases) while maintaining high recall. GPT-4o-mini reached 0.932 precision and 0.944 recall (F1 0.938), and Gemma 3 12B achieved an F1 of 0.848, demonstrating effective contextual reasoning. Some models, like Qwen3 8B and Aya Expanse 8B, prioritized recall (>0.99) at the cost of precision, over-classifying columns as sensitive. Others, including Gemma 2 9B and Qwen3 14B, offered a

better precision-recall balance, suitable for high-risk scenarios. Overall, the reflection step allowed nuanced sensitivity classification, significantly boosting precision with minimal recall loss and confirming the value of the detect-then-reflect approach.

6.1.3 End-to-End Evaluation (Detect-Then-Reflect). To assess the full pipeline performance of our *detect-then-reflect* mechanism, we reported the performance of LLMs in an end-to-end setting where both PII detection and contextual sensitivity reflection are applied. We compared three configurations: traditional baselines (i.e., treating all detected PII-type columns as sensitive), LLMs without reflection (i.e., treating all detected PII-type columns as sensitive), and LLMs with reflection (i.e., filtering detected PII columns based on table context).

System / Model	No Reflection			With Reflection		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Google DLP	0.531	0.628	0.576	–	–	–
Presidio	0.520	0.618	0.565	–	–	–
Ground-truth PII	0.527	**1.000**	0.690	–	–	–
GPT-4o-mini	0.856	0.639	0.732	**0.938**	0.632	0.755
Gemma 2 9B	0.740	0.819	0.778	0.800	0.792	0.796
Gemma 3 12B	0.487	0.941	0.642	0.753	0.806	0.779
Qwen3 8B	0.742	0.868	0.800	0.749	0.868	0.804
Qwen3 14B	0.565	<u>0.972</u>	0.714	0.732	0.941	<u>0.824</u>
Aya Expanse 8B	0.812	0.674	0.736	0.812	0.674	0.736
Qwen3 8B FT → GPT-4o-mini	–	–	–	<u>0.902</u>	0.861	**0.881**

Table 3. End-to-end sensitive-column classification. Each model is shown with and without the reflection step side by side. Reflection consistently improved precision and F1, with the best result (F1 = 0.881) from a hybrid pipeline: Qwen3 8B fine-tuned for detection combined with GPT-4o-mini for reflection.

In the no-reflection setup, LLMs marked any detected PII-type column as sensitive, leading to high recall but low precision; Gemma 3 12B and Qwen3 14B reached recall of 0.941 and 0.972 but precision of only 0.487 and 0.565, while GPT-4o-mini achieved higher precision (0.856) at lower recall (0.639). Introducing reflection using full-table context substantially improved precision across models, e.g., Qwen3 14B increased to 0.732 precision with 0.941 recall (F1 0.824) and GPT-4o-mini reached 0.938 precision. A modular pipeline combining Qwen3 8B (detection) with GPT-4o-mini (reflection) achieved F1 0.881. Overall, the *detect-then-reflect* strategy corrects the naive assumption that all PII-type columns are sensitive, boosting precision while maintaining recall, especially for ambiguous cases like `GENERIC_IDENTIFIER` or `ORGANIZATION_NAME`.

6.1.4 Reflection Only (Without Detection Input). When models were tasked with identifying sensitive columns using only the full table context (no PII labels), performance remained strong but recall dropped compared to the full pipeline. Qwen3 14B achieved an F1 of 0.819 (Precision=0.761, Recall=0.885), underscoring the value of the initial detection step for focusing the model's attention.

Model	Precision	Recall	F1
GPT-4o-mini	**0.964**	0.562	0.711
Gemma 2 9B	<u>0.878</u>	0.750	<u>0.809</u>
Qwen3 8B	0.735	<u>0.781</u>	0.758
Qwen3 14B	0.761	**0.885**	**0.819**
Aya Expanse 8B	0.815	0.674	0.738

Table 4. Reflection-only performance (without prior detection step) for identifying contextually sensitive columns. Evaluated across models using precision, recall, and F1 scores.

As shown in Table 4, all evaluated LLMs could identify sensitive columns using only table context, without prior PII-type labels. Qwen3 14B achieved the highest F1 (0.819) with a balanced precision (0.761) and recall (0.885), followed by Gemma 2 9B (F1 0.809). GPT-4o-mini excelled in precision (0.964) but had lower recall (0.562), while Qwen3 8B and Aya Expanse 8B offered moderate trade-offs (precision and recall ~0.7-0.8). These results indicate that LLMs can reason about sensitivity from context alone, but compared to the full *detect-then-reflect* setup (Table 3), recall is lower, highlighting the value of an initial detection pass to guide focus on likely PII columns.

6.2 Domain Contextualization (*Retrieve-Then-Detect*)

We evaluated the Retrieve-Then-Detect mechanism on a combined set of synthetic and real humanitarian tables (Table 5). Without external domain knowledge (ISP retrieval), all models exhibited a highly conservative bias, achieving perfect recall but low precision (<0.57). Integrating ISP retrieval into the prompt consistently improved precision. GPT-4o-mini's precision increased from 0.474 to 0.692 (F1=0.818), and Qwen3 14B's from 0.562 to 0.643 (F1=0.783), both maintaining perfect recall. Qualitative analysis revealed that the retrieval mechanism enabled models to produce interpretable justifications by citing specific clauses from the ISP documents, thereby enhancing the transparency and auditability of the sensitivity assessments.

System / Model	No Domain Knowledge			With Domain Knowledge		
	Prec.	Rec.	F1	Prec.	Rec.	F1
All-tables-sensitive (baseline)	0.375	1.000	0.545	–	–	–
GPT-4o-mini	0.474	1.000	0.643	0.692	1.000	**0.818**
Gemma 2 9B	0.375	1.000	0.545	0.429	1.000	0.600
Gemma 3 12B	0.529	1.000	0.692	0.500	1.000	0.667
Qwen3 8B	0.562	1.000	0.720	**0.778**	0.778	0.778
Qwen3 14B	0.562	1.000	0.720	0.643	1.000	0.783
Aya Expanse 8B	0.450	1.000	0.621	0.500	1.000	0.667

Table 5. Table-level classification performance on HDX and synthetic sensitive data with and without ISP-based retrieval. ISP retrieval improved precision while keeping recall near-perfect.

7 Conclusion

This paper explored how LLMs can detect sensitive information in tabular data by moving beyond fixed PII categories toward context-aware interpretations. Sensitive data was reconceptualized as context-dependent, operationalized through two mechanisms: *detect-then-reflect* for type contextualization and *retrieve-then-detect* for domain contextualization.

The first mechanism separated PII detection from contextual reflection, reducing false positives while maintaining recall. The second incorporated retrieval of country-specific Information Sharing Protocols (ISPs), enabling LLMs to distinguish between structurally similar tables with different sensitivities. Both mechanisms were evaluated on synthetic and real humanitarian datasets, benchmarking a range of open-source and proprietary models.

Results showed LLMs outperform rule-based systems in PII-type detection, with fine-tuned models such as Qwen3 8B FT achieving F1 > 0.94. Contextual cues like column names and sample values were critical for accuracy, while reflection improved precision (e.g., GPT-4o-mini and Qwen3 14B raising precision from <0.60 to >0.90). For domain-level classification, ISP retrieval enhanced precision without sacrificing recall, with models referencing policy text to justify predictions.

Overall, this work demonstrates that contextualization substantially improves sensitive data detection and positions LLMs as effective tools for data auditing. The proposed framework advances data-centric machine learning by aligning automated detection with local context and governance requirements.

References

Akhtar, M., O. Benjelloun, C. Conforti, L. Foschini, P. Gijsbers, J. Giner-Miguelez, S. Goswami, N. Jain, M. Karamousadakis, S. Krishna, M. Kuchnik, S. Lesage, Q. Lhoest, P. Marcenac, M. Maskey, P. Mattson, L. Oala, H. Oderinwale, P. Ruyssen, T. Santos, R. Shinde, E. Simperl, A. Suresh, G. Thomas, S. Tykhonov, J. Vanschoren, S. Varma, J. van der Velde, S. Vogler, C.-J. Wu, and L. Zhang (2024). Croissant: A metadata format for ml-ready datasets. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), *Advances in Neural Information Processing Systems*, Volume 37, pp. 82133–82148. Curran Associates, Inc.

Borgman, C. L. and P. Groth (2025). From data creator to data reuser: Distance matters. *Harvard Data Science Review*.

Brickley, D., M. Burgess, and N. Noy (2019). Google dataset search: Building a search engine for datasets in an open web ecosystem. In *The world wide web conference*, pp. 1365–1375.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei (2020). Language models are few-shot learners.

Carlini, N., F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. (2021). Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650.

Centers for Medicare & Medicaid Services (1996). The Health Insurance Portability and Accountability Act of 1996 (HIPAA). Online at http://www.cms.hhs.gov/hipaa/.

Cheng, J., M. Marone, O. Weller, D. Lawrie, D. Khashabi, and B. V. Durme (2024). Dated data: Tracing knowledge cutoffs in large language models.

European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

Fang, X., W. Xu, F. A. Tan, J. Zhang, Z. Hu, Y. Qi, S. Nickleach, D. Socolinsky, S. Sengamedu, and C. Faloutsos (2024). Large language models(llms) on tabular data: Prediction, generation, and understanding – a survey.

Google (2018). Data loss prevention.

Google Cloud (2018, March). Cloud Data Loss Prevention (DLP) API – General Availability. Web documentation. Accessed: 2025-07-22.

Hulsebos, M., Ç. Demiralp, and P. Groth (2023). Gittables: A large-scale corpus of relational tables. *Proceedings of the ACM on Management of Data 1*(1), 1–17.

Hulsebos, M., K. Hu, M. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, c. Demiralp, and C. Hidalgo (2019a). Sherlock: A deep learning approach to semantic data type detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.

Hulsebos, M., K. Hu, M. Bakker, E. Zgraggen, A. Satyanarayan, T. Kraska, c. Demiralp, and C. Hidalgo (2019b). Sherlock: A deep learning approach to semantic data type detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.

Ji, X., P. Glenn, A. G. Parameswaran, and M. Hulsebos (2025). Target: Benchmarking table retrieval for generative tasks. *arXiv preprint arXiv:2505.11545*.

Kober, M., J. Samhi, S. Arzt, T. F. Bissyandé, and J. Klein (2023). Sensitive and personal data: What exactly are you talking about? In *2023 IEEE/ACM 10th International Conference on Mobile Software Engineering and Systems (MOBILESoft)*, pp. 70–74. IEEE.

Kužina, V., A.-M. Petric, M. Barišić, and A. Jović (2023). Cassed: Context-based approach for structured sensitive data detection. *Expert Systems with Applications 223*, 119924.

Lukas, N., A. Salem, R. Sim, S. Tople, L. Wutschitz, and S. Zanella-Béguelin (2023). Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy*

(SP), pp. 346–363. IEEE.

Malkin, N. (2022). Contextual integrity, explained: A more usable privacy definition. *IEEE Security & Privacy 21*(1), 58–65.

Malkin, N. (2023). Contextual integrity, explained: A more usable privacy definition. *IEEE Security Privacy 21*(1), 58–65.

Microsoft (2018). Microsoft presidio.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Wash. L. Rev. 79*, 119.

OCHA, U. (2025). Data responsibility guidelines 2025.

OpenAI, :, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Mądry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A. T. Passos, A. Kirillov, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes, A. Tootoochian, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codispoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko, A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guarraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C. J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, D. Sherburn, D. Kappler, D. Levin, D. Levy, D. Carr, D. Farhi, D. Mely, D. Robinson, D. Sasaki, D. Jin, D. Valladares, D. Tsipras, D. Li, D. P. Nguyen, D. Findlay, E. Oiwoh, E. Wong, E. Asdar, E. Proehl, E. Yang, E. Antonow, E. Kramer, E. Peterson, E. Sigler, E. Wallace, E. Brevdo, E. Mays, F. Khorasani, F. P. Such, F. Raso, F. Zhang, F. von Lohmann, F. Sulit, G. Goh, G. Oden, G. Salmon, G. Starace, G. Brockman, H. Salman, H. Bao, H. Hu, H. Wong, H. Wang, H. Schmidt, H. Whitney, H. Jun, H. Kirchner, H. P. de Oliveira Pinto, H. Ren, H. Chang, H. W. Chung, I. Kivlichan, I. O'Connell, I. O'Connell, I. Osband, I. Silber, I. Sohl, I. Okuyucu, I. Lan, I. Kostrikov, I. Sutskever, I. Kanitscheider, I. Gulrajani, J. Coxon, J. Menick, J. Pachocki, J. Aung, J. Betker, J. Crooks, J. Lennon, J. Kiros, J. Leike, J. Park, J. Kwon, J. Phang, J. Teplitz, J. Wei, J. Wolfe, J. Chen, J. Harris, J. Varavva, J. G. Lee, J. Shieh, J. Lin, J. Yu, J. Weng, J. Tang, J. Yu, J. Jang, J. Q. Candela, J. Beutler, J. Landers, J. Parish, J. Heidecke, J. Schulman, J. Lachman, J. McKay, J. Uesato, J. Ward, J. W. Kim, J. Huizinga, J. Sitkin, J. Kraaijeveld, J. Gross, J. Kaplan, J. Snyder, J. Achiam, J. Jiao, J. Lee, J. Zhuang, J. Harriman, K. Fricke, K. Hayashi, K. Singhal, K. Shi, K. Karthik, K. Wood, K. Rimbach, K. Hsu, K. Nguyen, K. Gu-Lemberg, K. Button, K. Liu, K. Howe, K. Muthukumar, K. Luther, L. Ahmad, L. Kai, L. Itow, L. Workman, L. Pathak, L. Chen, L. Jing, L. Guy, L. Fedus, L. Zhou, L. Mamitsuka, L. Weng, L. McCallum, L. Held, L. Ouyang, L. Feuvrier, L. Zhang, L. Kondraciuk, L. Kaiser, L. Hewitt, L. Metz, L. Doshi, M. Aflak, M. Simens, M. Boyd, M. Thompson, M. Dukhan, M. Chen, M. Gray, M. Hudnall, M. Zhang, M. Aljubeh, M. Litwin, M. Zeng, M. Johnson, M. Shetty, M. Gupta, M. Shah, M. Yatbaz, M. J. Yang, M. Zhong, M. Glaese, M. Chen, M. Janner, M. Lampe, M. Petrov, M. Wu, M. Wang, M. Fradin, M. Pokrass, M. Castro, M. O. T. de Castro, M. Pavlov, M. Brundage, M. Wang, M. Khan, M. Murati, M. Bavarian, M. Lin, M. Yesildal, N. Soto, N. Gimelshein, N. Cone, N. Staudacher, N. Summers, N. LaFontaine, N. Chowdhury, N. Ryder, N. Stathas, N. Turley, N. Tezak, N. Felix, N. Kudige, N. Keskar, N. Deutsch, N. Bundick, N. Puckett, O. Nachum, O. Okelola, O. Boiko, O. Murk, O. Jaffe, O. Watkins, O. Godement, O. Campbell-Moore, P. Chao, P. McMillan, P. Belov, P. Su, P. Bak, P. Bakkum, P. Deng, P. Dolan, P. Hoeschele, P. Welinder, P. Tillet, P. Pronin, P. Tillet, P. Dhariwal, Q. Yuan, R. Dias, R. Lim, R. Arora, R. Troll, R. Lin, R. G. Lopes, R. Puri, R. Miyara, R. Leike, R. Gaubert, R. Zamani, R. Wang, R. Donnelly, R. Honsby, R. Smith, R. Sahai, R. Ramchandani, R. Huet, R. Carmichael, R. Zellers, R. Chen, R. Chen, R. Nigmatullin, R. Cheu, S. Jain, S. Altman, S. Schoenholz, S. Toizer, S. Miserendino, S. Agarwal, S. Culver, S. Ethersmith, S. Gray, S. Grove, S. Metzger, S. Hermani, S. Jain, S. Zhao, S. Wu, S. Jomoto, S. Wu, Shuaiqi, Xia, S. Phene, S. Papay, S. Narayanan, S. Coffey, S. Lee, S. Hall, S. Balaji, T. Broda, T. Stramer, T. Xu, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Cunninghman, T. Degry, T. Dimson, T. Raoux, T. Shadwell, T. Zheng, T. Underwood, T. Markov, T. Sherbakov, T. Rubin, T. Stasi, T. Kaftan, T. Heywood, T. Peterson, T. Walters, T. Eloundou, V. Qi, V. Moeller, V. Monaco, V. Kuo,

V. Fomenko, W. Chang, W. Zheng, W. Zhou, W. Manassra, W. Sheu, W. Zaremba, Y. Patil, Y. Qian, Y. Kim, Y. Cheng, Y. Zhang, Y. He, Y. Zhang, Y. Jin, Y. Dai, and Y. Malkov (2024). Gpt-4o system card.

Quinn, P. and G. Malgieri (2021). The difficulty of defining sensitive data—the concept of sensitive data in the eu data protection framework. *German Law Journal 22*(8), 1583–1612.

Raman, V. (2001). Potter's wheel: An interactive data cleaning system. In *VLDB*, Volume 1, pp. 381–390.

Shen, Y., Z. Ji, J. Lin, and K. R. Koedginer (2025). Enhancing the de-identification of personally identifiable information in educational data.

Shvartzshnaider, Y., N. Apthorpe, N. Feamster, and H. Nissenbaum (2019). Going against the (appropriate) flow: A contextual integrity approach to privacy policy analysis. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Volume 7, pp. 162–170.

Subramani, N., S. Luccioni, J. Dodge, and M. Mitchell (2023). Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pp. 208–220.

Team, G. (2025). Gemma 3.

Team, G., M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. L. Lan, S. Jerome, A. Tsitsulin, N. Vieillard, P. Stanczyk, S. Girgin, N. Momchev, M. Hoffman, S. Thakoor, J.-B. Grill, B. Neyshabur, O. Bachem, A. Walton, A. Severyn, A. Parrish, A. Ahmad, A. Hutchison, A. Abdagic, A. Carl, A. Shen, A. Brock, A. Coenen, A. Laforge, A. Paterson, B. Bastian, B. Piot, B. Wu, B. Royal, C. Chen, C. Kumar, C. Perry, C. Welty, C. A. Choquette-Choo, D. Sinopalnikov, D. Weinberger, D. Vijaykumar, D. Rogozińska, D. Herbison, E. Bandy, E. Wang, E. Noland, E. Moreira, E. Senter, E. Eltyshev, F. Visin, G. Rasskin, G. Wei, G. Cameron, G. Martins, H. Hashemi, H. Klimczak-Plucińska, H. Batra, H. Dhand, I. Nardini, J. Mein, J. Zhou, J. Svensson, J. Stanway, J. Chan, J. P. Zhou, J. Carrasqueira, J. Iljazi, J. Becker, J. Fernandez, J. van Amersfoort, J. Gordon, J. Lipschultz, J. Newlan, J. yeong Ji, K. Mohamed, K. Badola, K. Black, K. Millican, K. McDonell, K. Nguyen, K. Sodhia, K. Greene, L. L. Sjoesund, L. Usui, L. Sifre, L. Heuermann, L. Lago, L. McNealus, L. B. Soares, L. Kilpatrick, L. Dixon, L. Martins, M. Reid, M. Singh, M. Iverson, M. Görner, M. Velloso, M. Wirth, M. Davidow, M. Miller, M. Rahtz, M. Watson, M. Risdal, M. Kazemi, M. Moynihan, M. Zhang, M. Kahng, M. Park, M. Rahman, M. Khatwani, N. Dao, N. Bardoliwalla, N. Devanathan, N. Dumai, N. Chauhan, O. Wahltinez, P. Botarda, P. Barnes, P. Barham, P. Michel, P. Jin, P. Georgiev, P. Culliton, P. Kuppala, R. Comanescu, R. Merhej, R. Jana, R. A. Rokni, R. Agarwal, R. Mullins, S. Saadat, S. M. Carthy, S. Cogan, S. Perrin, S. M. R. Arnold, S. Krause, S. Dai, S. Garg, S. Sheth, S. Ronstrom, S. Chan, T. Jordan, T. Yu, T. Eccles, T. Hennigan, T. Kocisky, T. Doshi, V. Jain, V. Yadav, V. Meshram, V. Dharmadhikari, W. Barkley, W. Wei, W. Ye, W. Han, W. Kwon, X. Xu, Z. Shen, Z. Gong, Z. Wei, V. Cotruta, P. Kirk, A. Rao, M. Giang, L. Peran, T. Warkentin, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, D. Sculley, J. Banks, A. Dragan, S. Petrov, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, S. Borgeaud, N. Fiedel, A. Joulin, K. Kenealy, R. Dadashi, and A. Andreev (2024). Gemma 2: Improving open language models at a practical size.

United Nations (2014). The humanitarian data exchange. https://data.humdata.org/. Accessed: 2015-05-31.

Wang, Z., Y. Pang, and Y. Lin (2023). Large language models are zero-shot text classifiers.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data 3*(1), 1–9.

Worth, S., B. Snaith, A. Das, G. Thuermer, and E. Simperl (2024). Ai data transparency: an exploration through the lens of ai incidents. *arXiv preprint arXiv:2409.03307*.

Yang, A., A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu (2025). Qwen3 technical report.

Yang, J., X. Zhang, K. Liang, and Y. Liu (2023). Exploring the application of large language models in detecting and protecting personally identifiable information in archival data: A comprehensive study*. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2116–2123.

Yang, X., W. Liang, and J. Zou (2024). Navigating dataset documentations in ai: A large-scale analysis of dataset cards on hugging face.

Zhang, J., Y. Bi, M. Cheng, J. Liu, K. Ren, Q. Sun, Y. Wu, Y. Cao, R. C. Fernandez, H. Xu, et al. (2024). A survey on data markets. *arXiv preprint arXiv:2411.07267*.

Appendix A PII Entity Taxonomy

The following table lists all Personally Identifiable Information (PII) entity types used in this study:

Entity Name	
CREDIT_CARD_NUMBER	DATE_OF_BIRTH
DISABILITY_GROUP	EDUCATION_LEVEL
EMAIL_ADDRESS	ETHNIC_GROUP
GENDER	GENERIC_IDENTIFIER
GEO_COORDINATES	IBAN_CODE
IP_ADDRESS	MARITAL_STATUS
MEDICAL_TERM	OCCUPATION
ORGANIZATION_NAME	PASSPORT
PERSON_NAME	PHONE_NUMBER
PROTECTION_GROUP	RELIGIOUS_GROUPS
SEXUALITY	SPOKEN_LANGUAGE
STREET_ADDRESS	SWIFT_CODE
URL	ZIPCODE

TABLE 6. List of PII entity types used during PII detection.

PII Entity	Sensitivity ratio	Ratio
DATE_OF_BIRTH	17/17	1.00
EMAIL_ADDRESS	30/31	0.97
ETHNIC_GROUP	17/18	0.94
GENDER	11/14	0.79
GENERIC_IDENTIFIER	26/216	0.12
GEO_COORDINATES	7/24	0.29
IP_ADDRESS	0/3	0.00
MARITAL_STATUS	1/1	1.00
OCCUPATION	3/5	0.60
ORGANIZATION_NAME	1/27	0.04
PERSON_NAME	92/93	0.99
PHONE_NUMBER	20/21	0.95
SPOKEN_LANGUAGE	1/1	1.00
STREET_ADDRESS	35/38	0.92
URL	1/11	0.09
ZIPCODE	26/27	0.96

TABLE 7. Overview of sensitive occurrence ratios per PII entity type. The ratio reflects the proportion of times an entity was labeled as sensitive out of its total occurrences across all tables. High ratios (e.g., for `DATE_OF_BIRTH` and `PERSON_NAME`) indicate consistently sensitive entities, while others (e.g., `GENERIC_IDENTIFIER`, `URL`) are often benign and context-dependent.

Presidio Entity	Mapped Category
AU_ABN	GENERIC_ID
AU_ACN	GENERIC_ID
AU_MEDICARE	GENERIC_ID
AU_TFN	GENERIC_ID
CREDIT_CARD	CREDIT_CARD_NUMBER
CRYPTO	GENERIC_ID
DATE_TIME	None
EMAIL_ADDRESS	EMAIL_ADDRESS
ES_NIE	GENERIC_ID
ES_NIF	GENERIC_ID
FI_PERSONAL_IDENTITY_CODE	GENERIC_ID
IBAN_CODE	GENERIC_ID
IN_AADHAAR	GENERIC_ID
IN_PAN	GENERIC_ID
IN_PASSPORT	GENERIC_ID
IN_VEHICLE_REGISTRATION	GENERIC_ID
IN_VOTER	GENERIC_ID
IP_ADDRESS	LOCATION
IT_DRIVER_LICENSE	GENERIC_ID
IT_FISCAL_CODE	GENERIC_ID
IT_IDENTITY_CARD	GENERIC_ID
IT_PASSPORT	GENERIC_ID
IT_VAT_CODE	GENERIC_ID
LOCATION	None
MEDICAL_LICENSE	GENERIC_ID
NRP	ETHNIC_GROUP
None	None
PERSON	PERSON_NAME
PHONE_NUMBER	PHONE_NUMBER
PL_PESEL	GENERIC_ID
SG_NRIC_FIN	GENERIC_ID
SG_UEN	GENERIC_ID
UK_NHS	GENERIC_ID
UK_NINO	GENERIC_ID
URL	URL
US_BANK_NUMBER	GENERIC_ID
US_DRIVER_LICENSE	GENERIC_ID
US_ITIN	GENERIC_ID
US_PASSPORT	GENERIC_ID
US_SSN	GENERIC_ID
None	None

TABLE 8. Mapping of Presidio PII entities to the unified taxonomy used in this study.

Appendix B Synthetic Data Generation

Category	Count	Percentage
None	331	33.10
GENERIC_ID	279	27.90
DATE_OF_BIRTH	50	5.00
PERSON_NAME	44	4.40
ETHNIC_GROUP	37	3.70
STREET_ADDRESS	34	3.40
PHONE_NUMBER	34	3.40
GENDER	30	3.00
EMAIL_ADDRESS	27	2.70
GEO_COORDINATES	25	2.50
URL	24	2.40
ORGANIZATION_NAME	23	2.30
ZIPCODE	21	2.10
OCCUPATION	12	1.20
MARITAL_STATUS	11	1.10
IMEI_HARDWARE_ID	5	0.50
SPOKEN_LANGUAGE	3	0.30
IP_ADDRESS	2	0.20
SWIFT_CODE	2	0.20
MEDICAL_TERM	1	0.10
PROTECTION_GROUP	1	0.10
CREDIT_CARD_NUMBER	1	0.10
IBAN_CODE	1	0.10
ICD9_CODE	1	0.10
EDUCATION_LEVEL	1	0.10

TABLE 9. Distribution of categories in the `output` column of the synthetic dataset for PII-type detection.

Appendix C Ablation Study Results

System / Model	Precision	Recall	F1
Google DLP weighted	0.713	0.324	0.298
Google DLP macro	0.409	0.322	0.296
Presidio weighted	0.198	0.245	0.208
Presidio macro	0.158	0.182	0.156
GPT-4o-mini weighted	0.578	0.366	0.422
GPT-4o-mini macro	0.735	0.503	0.555
GPT-4o-mini no records weighted	0.393	0.252	0.284
GPT-4o-mini no records macro	0.581	0.408	0.442
GPT-4o weighted	0.962	0.662	0.730
GPT-4o macro	0.702	0.664	0.668
Gemma 2 9B weighted	0.960	0.537	0.605
Gemma 2 9B macro	0.763	0.608	0.623
Gemma 2 9B FT weighted	0.947	0.775	0.833
Gemma 2 9B FT macro	0.748	0.687	0.703
Gemma 3 12B weighted	0.861	0.812	0.819
Gemma 3 12B macro	0.690	0.722	0.681
Qwen3 8B weighted	0.919	0.548	0.581
Qwen3 8B macro	0.723	0.618	0.627
Qwen3 8B FT weighted	0.968	0.932	0.941
Qwen3 8B FT macro	0.861	0.807	0.819
Qwen3 14B weighted	0.933	0.819	0.858
Qwen3 14B macro	0.720	0.735	0.724
Aya Expanse 8B weighted	0.931	0.393	0.448
Aya Expanse 8B macro	0.537	0.373	0.405

Table 10. Showing macro average as well. Performance of baseline systems and LLMs on PII entity detection in tabular data. Each system receives a column name and five unique sample values and predicts one of 27 PII types or "None". Reported are the weighted averages of precision, recall, and F1 score across all PII categories.

System / Model	Precision	Recall	F1
Google DLP	0.713	0.324	0.298
Presidio	0.198	0.245	0.208
GPT-4o-mini (k=5)	0.578	0.366	0.422
GPT-4o-mini (k=10)	0.577	0.400	0.448
Gemma 2 9B (k=5)	0.960	0.538	0.605
Gemma 2 9B (k=10)	0.949	0.572	0.642
Gemma 3 12B (k=5)	0.859	0.811	0.818
Gemma 3 12B (k=10)	0.882	0.812	0.830
Qwen3 8B (k=5)	0.919	0.549	0.582
Qwen3 8B (k=10)	0.909	0.563	0.582
Qwen3 14B (k=5)	0.931	0.819	0.857
Qwen3 14B (k=10)	0.911	0.819	0.841
Aya Expanse 8B (k=5)	0.931	0.394	0.448
Aya Expanse 8B (k=10)	0.920	0.413	0.456

Table 11. Ablation study of the performance of large language models (LLMs) on PII entity detection in tabular data. Each system receives a column name and five or ten unique sample values and predicts one of 27 PII types or "None". Reported are the weighted averages of precision, recall, and F1 score across all PII categories.