DataSIR: A Benchmark Dataset for Sensitive Information Recognition

Fan Mo, Bo Liu*, Yuan Fan, Kun Qin, Yizhou Zhao, Jinhe Zhou, Jia Sun, Jinfei Liu, Kui Ren
Zhejiang University
DBAPPSecurity Co., Ltd

Motivation

Original Data to Format Transformations

Data leakage prevention (DLP) technologies lag behind evolving evasion techniques. For example, various format transformations can be performed on the data, such as Unicode encoding, and then after leakage, reverse Unicode encoding can restore the original data.

Increasing Cost of A Data Leakage

The global average cost of a data leakage in 2024 rose to $4.88 million, an increase of nearly 10% from $4.45 million in 2023.

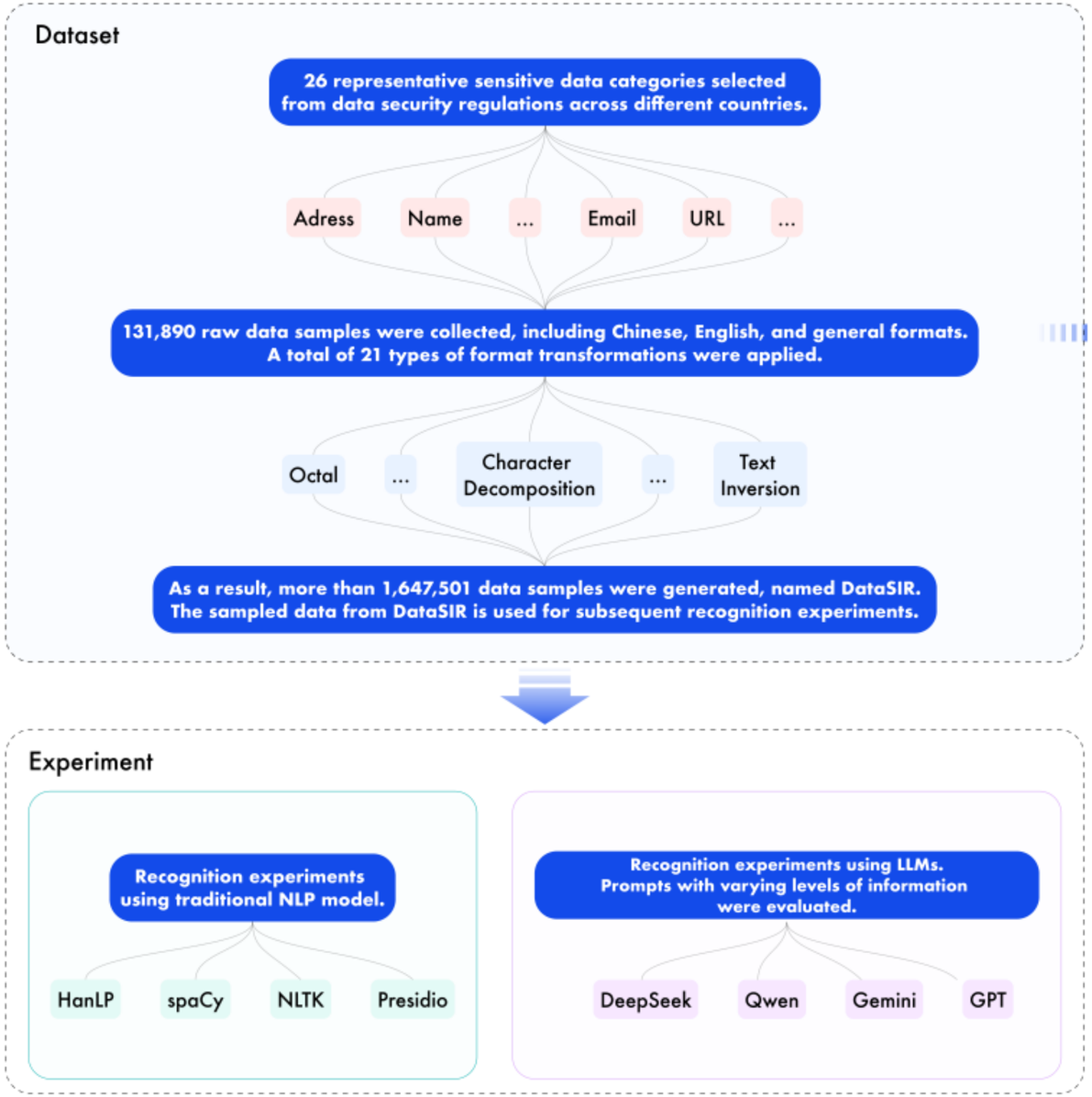
Lack of Datasets for Developing SIR Models

Current datasets lack comprehensive coverage of these adversarial transformations, limiting the evaluation of robust SIR systems.

What do we do?

We introduce DataSIR, a benchmark dataset specifically designed to evaluate SIR models on sensitive data subjected to diverse format transformations. We curate 26 sensitive data categories based on multiple international regulations, and collect 131,890 original samples correspondingly.

Through empirical analysis of real-world evasion tactics, we implement 21 format transformation methods, which are applied to the original samples, expanding the dataset to 1,647,501 samples to simulate adversarial scenarios.



Our Contributions

Multilingual and Rich-Regulations Coverage

To ensure consistency and broad applicability, 26 representative sensitive data categories were selected based on major international regulations (e.g., HIPAA, SOX, GDPR, CCPA, PIPL). And examples were provided in both Chinese and English.

Extensive Format Transformations

For each sensitive category, 21 transformation types (e.g., binary, octal, Morse code, insertion of digits or English words) are applied, resulting in 1,647,501 samples, which significantly enrich the diversity of sensitive data.

High-Quality Benchmark Dataset

The dataset’s quality was validated using various NLP and LLM methods and models, demonstrating strong differentiation capabilities across different categories and formats. It can serve as a robust benchmark for evaluating and developing future sensitive information recognition models.

The DataSIR Dataset - 26 Types of Data

Table 1: Overviews of DataSIR

Category	Covered Regulations	Language Involved	Original Count	Transformed Count	Total Count
Address	GDPR, PIPL, CCPA	Chinese/English	6000	72000	78000
Marital Status	GDPR, PIPL, CCPA	Chinese/English	8	104	112
Medical History	HIPAA, PIPL, CCPA	Chinese/English	6000	74838	80838
Name	GDPR, PIPL, CCPA	Chinese/English	6000	77607	83607
Nationality	GDPR, PIPL, CCPA	Chinese/English	482	6204	6686
Occupation	GDPR, PIPL, CCPA	Chinese/English	600	7542	8142
Organization	HIPAA, SOX, GDPR	Chinese/English	6000	73345	79345
Party	GDPR, PIPL, CCPA	Chinese/English	600	7402	8002
Religion	GDPR, PIPL, CCPA	Chinese/English	200	2569	2769
Date/Time	HIPAA, SOX	General	6000	48000	54000
Driver's License	GDPR, PIPL, CCPA	General	6000	66000	72000
Email	GDPR, PIPL, CCPA	General	6000	66000	72000
Personal ID	GDPR, PIPL, CCPA	General	6000	66000	72000
IMEI	GDPR, PIPL, CCPA	General	6000	84000	90000
IMSI	GDPR, PIPL, CCPA	General	6000	84000	90000
IPv4	GDPR, PIPL, CCPA	General	6000	66000	72000
IPv6	GDPR, PIPL, CCPA	General	6000	72000	78000
JDBC Connection String	GDPR, PIPL, CCPA	General	6000	66000	72000
Landline Number	HIPAA, CCPA	General	8000	96000	104000
MAC	GDPR, PIPL, CCPA	General	6000	72000	78000
MEID	GDPR, PIPL, CCPA	General	6000	66000	72000
Mobile Number	GDPR, PIPL, CCPA	General	8000	96000	104000
Passport	GDPR, PIPL, CCPA	General	6000	66000	72000
Postcode	GDPR, PIPL, CCPA	General	6000	66000	72000
Transaction Amount	GDPR, PIPL, CCPA, SOX	General	6000	48000	54000
URL	GDPR, PIPL, CCPA	General	6000	66000	72000

The DataSIR Dataset - 21 Format Transformations

- A. Binary
- B. Octal
- C. Hexadecimal
- D. ASCII encoding
- E. Unicode encoding
- F. UTF-8 encoding
- G. Base64 encoding
- H. URL encoding
- I. HTML entity encoding
- J. Morse encoding
- K. Braille encoding
- L. Nested encoding
- M. Acrostic poetry
- N. Character decomposition
- O. Text inversion
- P. Martian text
- Q. Simplified to traditional Chinese
- R. Numerical capitalization
- S. Inserting special characters
- T. Inserting Chinese characters
- U. Inserting English letters/numbers

Base-8 system using digits 0-7. (e.g., 616655990822147 → 6 1 6 6 5 5 11 11 0 10 2 2 1 4 7).

Converting byte streams obtained through UTF-8 encoding into 64 printablecharacters (A-Z, a-z, 0-9, +, /). (e.g., China → Q2hpbmE=).

Using combinations of short (·) and long (—·) signals to represent let-ters/numbers, separated by spaces between words. (e.g., China → -.-. .... .. -. .-).

Decomposing Chinese characters into components or strokes. (e.g., 功 = 工 + 力).

Reversing character sequence. (e.g., hello → olleh).

The DataSIR Dataset - Cross-Reference Table

- Not all sensitive categories can undergo all 21 format transformations. Some transformations are applicable to Chinese, others to English, some to numbers, and others to symbols.
- Each category is applicable to around 10 transformations, with a minimum of 8 and a maximum of 14.

Table 2: Sensitive Category - Format Transformation Cross-Reference Table

Category	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Address	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	✓	✓	✓
Marital Status	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
Medical History	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
Name	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
Nationality	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
Occupation	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
Organization	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
Party	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
Religion	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
Date/Time	×	×	×	×	✓	✓	✓	×	✓	×	×	✓	×	×	×	✓	×	✓	✓	×	×
Driver's License	×	×	×	✓	✓	✓	✓	×	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
Email	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	✓	×	×	✓	×	×
Personal ID	×	×	×	✓	✓	✓	✓	×	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
IMEI	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
IMSI	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
IPv4	×	×	×	✓	✓	✓	✓	×	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
IPv6	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
JDBC Connection string	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	✓	×	×	✓	×	×
Landline Number	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
MAC	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
MEID	×	×	×	✓	✓	✓	✓	×	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
Mobile Number	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
Passport	×	×	×	✓	✓	✓	✓	×	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
Postcode	×	×	×	✓	✓	✓	✓	×	✓	✓	✓	✓	×	×	×	✓	×	✓	✓	×	×
Transaction Amount	×	×	×	×	✓	✓	✓	×	✓	×	×	✓	×	×	×	✓	×	✓	✓	×	×
URL	×	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	✓	×	×	✓	×	×

Experiment Results- NLP vs LLMs

- The commonly used NLP models and tools in traditional data security solutions perform poorly in defending against advanced data leaks.

Table 3: Comparison of LRAcc for NLP Model Based Tools

Tool	Labels Count	List of Recognizable Labels	Original	Transformed	Overall
HanLP	8	Landline, Mobile Number, Date/Time, Postal Code, Amount, Address, Name, Organization	13.71%	**4.15%**	4.91%
spaCy	8	Date/Time, Amount, Nationality, Address, Name, Party Affiliation, Organization, Religious	13.29%	2.40%	2.98%
NLTK	3	Address, Organization, Name	2.59%	0.39%	0.56%
Presidio	12	IPv4, URL, Landline, Mobile Number, Date/Time, Email, Nationality, Address, Name, Party	**23.71%**	3.31%	**4.93%**

- If traditional data security solutions can integrate LLMs, the effectiveness of defending against data leaks would improve significantly and has the potential for further enhancement.
- As the information content in the prompts increases, the LRAcc also increases.

Table 4: Comparison of LRAcc for LLMs with Different Prompts

Prompts	DeepSeek LRAcc	Qwen LRAcc	Gemini LRAcc	GPT LRAcc
no label info, no format info	4.18%	5.68%	4.46%	**6.65%**
with label info, no format info	47.90%	47.55%	53.91%	**55.79%**
with label info, with format info	54.37%	55.97%	**65.04%**	64.30%

Experiment Results - Comparison of Results for Gemini with Different Format Transformation

Table 5: Comparison of Results for Gemini with Different Format Transformation

Type	LRAcc (%)	DRAcc (%)
Binary	18.00	98.00
Octal	18.00	98.00
Hexadecimal	16.00	0.00
ASCII encoding	69.57	95.74
Unicode encoding	71.39	97.17
UTF-8 encoding	72.43	95.53
Base64 encoding	59.02	66.47
URL encoding	86.02	97.49
HTML entity encoding	70.64	94.78
Morse encoding	63.37	69.77
Braille encoding	52.71	46.51
Nested encoding	57.68	60.21
Acrostic poetry	71.85	76.30
Character decomposition	66.35	61.54
Text inversion	68.57	57.96
Martian text	61.25	58.27
Simplified to traditional Chinese	74.04	50.96
Numerical capitalization	47.86	78.35
Inserting special characters	66.02	68.71
Inserting Chinese characters	80.14	85.82
Inserting English letters/numbers	65.38	58.65
All Above Format Transformed Data	64.39	75.26
Original data	72.58	95.08

1. The LRAcc and DRAcc of all format-transformed data are lower than those of the original data, which indicates that it is more difficult to recognize and restore data after format transformed.

2. Gemini's recognition of URL-encoded data is the best, as URL encoding only involves transforming Chinese characters and some symbols, making it relatively easy for LLMs to restore the original data and significantly enhancing the recognition of sensitive categories.

Experiment Results - Comparison of Results for Gemini with Different Sensitive Categories

Table 6: Comparison of Results for Gemini with Different Sensitive Categories

Category	Precision (%)	Recall (%)	F1-score (%)	DRAcc (%)
Address	62.65	99.08	76.76	61.85
Marital Status	90.80	95.62	93.15	89.69
Medical History	99.57	69.85	82.11	62.99
Name	65.11	92.51	76.43	76.08
Nationality	95.15	56.48	70.89	80.69
Occupation	97.65	62.09	75.91	71.64
Organization	40.89	78.05	53.67	65.85
Party	87.93	30.63	45.43	76.88
Religion	99.07	30.72	46.90	65.80
Date/Time	96.90	93.59	95.22	84.62
Driver's License	16.67	0.67	1.28	75.00
Email	91.46	96.66	93.98	63.21
Personal ID	25.79	65.67	37.03	74.67
IMEI	26.03	21.87	23.77	83.20
IMSI	83.33	6.67	12.35	86.93
IPv4	95.62	94.67	95.14	87.00
IPv6	98.95	87.38	92.81	69.54
JDBC Connection string	97.39	99.67	98.52	69.67
Landline Number	62.69	74.46	68.07	76.92
MAC	62.77	89.23	73.70	76.31
MEID	68.29	18.73	29.40	65.22
Mobile Number	27.47	80.31	40.94	78.77
Passport	0.00	0.00	0.00	80.67
Postcode	73.50	77.67	75.53	93.00
Transaction Amount	71.72	31.56	43.83	64.89
URL	95.81	99.00	97.38	73.33

1. The model achieves F1-score above 90% in categories such as URL, JDBC Connection String, IPv4 and IPv6 Address, Email, and Date/Time. This indicates that even after format transformations, these categories of sensitive information with stable or distinctive formatting patterns can still be effectively recognized.

2. In contrast, categories such as Personal ID, Passport, Driver's License, and Transaction Amount generally yield F1-score below 40%. This performance degradation is mainly due to the fact that recognizing these categories of sensitive data depends more on contextual semantics and discourse cues.

THANK YOU

kaggle

Github