SPY: Enhancing Privacy with Synthetic PII Detection Dataset

Maksim Savkin[4,1,*] Timur Ionov[2,3,*] Vasily Konovalov[4]

[1]MIPT [2]MWS AI [3]ITMO University [4]AIRI

Introduction

Introducing the SPY Dataset — a novel, fully synthetic dataset for detecting Personally Identifiable Information (PII). SPY highlights the important difference between traditional named entities and personal data, which standard Named Entity Recognition (NER) models often treat as the same. The Table 1 illustrates this distinction.

Traditional named entities vs PII data

a) Apple technical support for education customers: 1-800-800-2775.
Satya Nadella is CEO of Microsoft Corp.

b) Lucy Cechtelar lives at 426 Jordy Lodge Cartwrightshire, SC 88120-6700.

Table 1. Examples of **a)** NER entities; **b)** PII entities. All examples of personal information provided are generated using the Faker library [2].

- **Problem:** Most public datasets [1] and tools [4] detect entity types (e.g., name, email) rather than actual personal data.
- **Solution:** SPY generates fully synthetic texts with realistic fake PII, enabling safe and effective training and evaluation.

Dataset construction

Step 1) Generate diverse texts for specified domain using fake biographies

Step 1) Look through the personality of the text author and pretend to be that person.
occupation: *<generated-occupation>*, personality: *<generated-personality>*
Step 2) Use the following instructions to generate a text:*<domain-instructions>*
Requirements:
- At any circumstance do not include any personal information in generated text.
Respond only with generated text with no commentary. Here goes your text:

↓

Step 2) Iteratively insert new PII placeholders to increase entity concentration

Text: ... **Task:** You are an author of the above Text. Your task is to add new placeholders in the Text from the list below. You will be penalized for mentioning any placeholders other than what is listed below! **Here is the list of placeholders representing your personal information:**
<author_personal_name> - A full or partial name of the text author, <...> - ...
Requirements: 1) Do NOT change existing placeholders 2) Distribute placeholders evenly throughout your text, do not stack them all in one place 3) New text must be more entity-dense than the previous one
Respond only with updated text with no commentary. Here goes an updated text:

Repeat × n

↓

Step 3) Replace placeholders with synthetic entities

The paper was submitted by <author_personal_name> and reviewed by three anonymous reviewers.

↓ Synthetic Entities Generator ↓

The paper was submitted by Emily Johnson and reviewed by three anonymous reviewers.

↓

Step 4) Add traditional named entities to strengthen the contrast between PII and general NER

Text: ... **Task:** You are given a Text, which contains author's personal information. Your task is to add new entities, which are not related to the text author. Generate entities using the following classes: name, email, username, phone number, url, address, identifier.
Requirements: 1) At any circumstance DO NOT change author's personal information in the above text 2) Newly generated entities should not disclose the personal information of the author of the text
Respond only with updated text with no commentary. Here goes an updated text:

Data Analysis

- **PII Density Control**: The update mechanism boosts entity-rich text generation. Table 2, column *PII update 2*, shows a steady rise in entity frequency.
- **Entity Balance**: All entity types are evenly represented, each comprising approximately 12–15% of the total entities, see Figure 1.
- **Control over traditional named entities**: The pipeline allows to control the inclusion of non-PII entities, such as public names or locations.

Entity	Legal questions			Medical Consultations		
	PII update		Add non-PII	PII update		Add non-PII
	#1	#2		#1	#2	
Name	0.6	1.1 (+0.5)	0.9 (+0.3)	0.7	1.1 (+0.4)	1.0 (+0.3)
Email	1.0	1.2 (+0.1)	0.9 (-0.2)	1.0	1.1 (+0.1)	0.9 (-0.1)
Username	0.9	1.1 (+0.2)	1.3 (+0.4)	0.8	1.2 (+0.4)	1.3 (+0.5)
Phone	0.9	1.1 (+0.2)	0.8 (-0.1)	0.9	1.1 (+0.2)	0.9 (+0.0)
URL	1.1	1.3 (+0.3)	0.9 (-0.2)	1.0	1.3 (+0.3)	0.9 (-0.2)
Address	0.7	1.2 (+0.5)	0.9 (+0.2)	0.7	1.3 (+0.6)	1.1 (+0.3)
ID	0.4	1.0 (+0.6)	0.7 (+0.3)	0.5	1.1 (+0.5)	0.9 (+0.4)
Average	0.8	1.1 (+0.4)	0.9 (+0.1)	0.8	1.2 (+0.4)	1.0 (+0.2)

Table 2. Average number of PII entities detected in texts generated by SPY prompting pipeline. Each entity type is counted separately. *PII update #{k}* refers to the average number of PII entities in texts after k iterative updates of PII placeholders; *Add non-PII* represents the average number of PII entities in texts that have completed all stages of the pipeline.

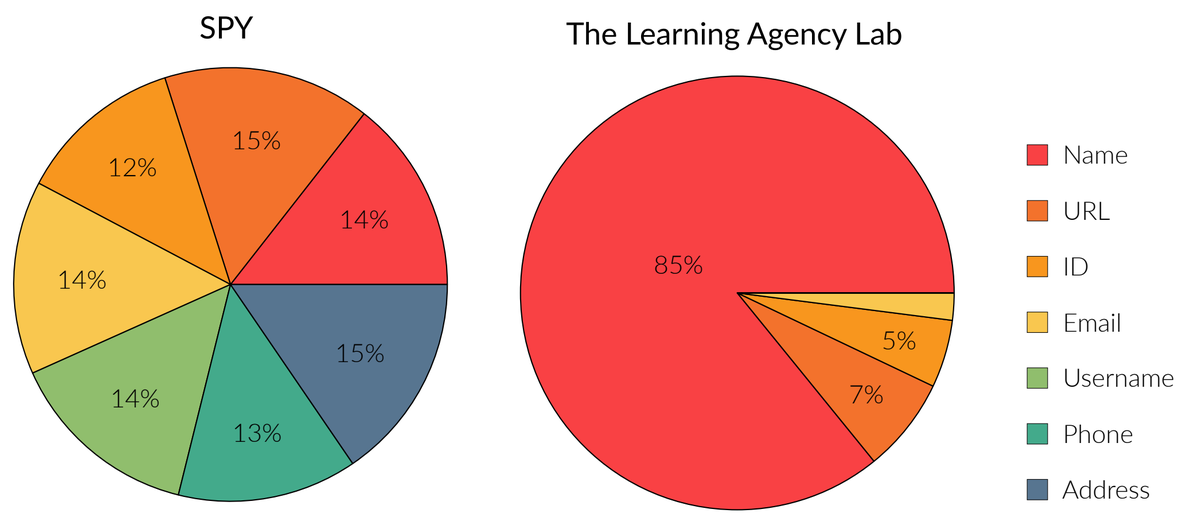

Figure 1. Side-by-side comparison of entity type distributions for legal-domain in SPY and the PII detection dataset from "Learning Agency Lab" [3].

Experimental Results

We compare three baselines for PII detection in the Table 3:

- **Presidio** (rule-based, regexp and NER): High Recall but low Precision; misclassifies many standard NER entities as PII.
- **Llama-3-70B** (zero-shot): Outperforms Presidio by better distinguishing between PII and generic named entities, but still struggles with boundary precision.
- **DeBERTa (our)** (fine-tuned): Achieves the highest F1-score; confirms the benefit of training specifically for PII detection.

Entity	Metric	Medical Consultations			Legal Questions		
		Llama-3-70B	Presidio	DeBERTa (our)	Llama-3-70B	Presidio	DeBERTa (our)
Name	P	73.0	17.1	**86.9**	64.7	17.9	**87.4**
	R	62.9	80.4	**88.7**	68.9	79.4	**93.2**
	F1	67.6	28.2	**87.8**	66.7	29.2	**90.2**
Email	P	92.7	37.6	**97.6**	91.8	33.7	**92.1**
	R	90.9	92.2	**99.5**	88.5	91.8	**99.1**
	F1	91.8	53.4	**98.5**	90.1	49.3	**95.5**
Username	P	68.8	-	**92.1**	66.1	-	**90.3**
	R	70.4	-	**95.4**	59.7	-	**98.0**
	F1	69.6	-	**93.8**	62.7	-	**94.0**
URL	P	83.6	6.9	**97.5**	84.5	7.9	**94.4**
	R	91.9	19.4	**98.9**	92.5	21.3	**99.0**
	F1	87.5	10.2	**98.2**	88.3	11.5	**96.7**
ID	P	91.7	26.1	**96.7**	91.9	20.6	**93.0**
	R	75.1	38.9	**98.3**	62.2	34.4	**96.6**
	F1	82.6	31.2	**97.5**	74.2	25.8	**94.8**
Phone	P	89.8	37.4	**93.3**	85.7	34.1	**87.5**
	R	90.0	65.5	**96.9**	92.8	68.1	**98.7**
	F1	89.9	47.6	**95.0**	89.1	45.4	**92.8**
Address	P	**96.2**	-	89.3	**93.7**	-	88.3
	R	90.4	-	**95.1**	81.3	-	**94.5**
	F1	**93.2**	-	92.1	87.1	-	**91.3**

Table 3. Performance metrics. **Presidio** [4] is a Microsoft SDK for fast PII detection using NER, regex, rule-based logic. **LLaMA-3-70B-zero-shot** is a zero-shot prompted LLM for PII task. **DeBERTa** is a model cross-validated on different domains of SPY dataset. Blanks mean that entity class is not supported.

References

[1] Ai4Privacy. *Open PII Masking 500k Dataset*. Hugging Face, 2025.
[2] Daniele Faraglia. *Faker*. 2014.
[3] Langdon Holmes et al. *The Learning Agency Lab - PII Data Detection*. Kaggle. 2024.
[4] Microsoft. *Presidio*. 2021.

Savkin@airi.net