Large Language Model–Based Synthetic Data Generation for Enhanced Personal Information Classification

Anonymous ACL submission

Abstract

In Personally Identifiable Information (PII) detection, training high-accuracy classifiers requires substantial annotated data, yet collecting real data is challenging due to privacy constraints. This study proposes a method to address this challenge through synthetic data generation using Large Language Models (LLMs). Our approach employs a simple configuration that provides few-shot examples to high-performance LLMs, eliminating the need for complex pipelines or domain-specific rule design. In experiments on the SPeDaC1 benchmark, we generated 9,000 synthetic samples from only 100 seed examples and achieved an accuracy of 0.927. This represents a dramatic performance improvement compared to training on small real datasets alone (20–100 examples, accuracy 0.490), demonstrating that LLM-based synthetic data generation is an effective means of overcoming the limitations of small-data learning. Our method enables rapid deployment to new domains and low-resource languages, contributing to the balance between privacy protection and AI development.

1 Introduction

As global data distribution advances, the proper management of Personally Identifiable Information (PII) serves as a critical foundation for maintaining social trust.

The European Commission protects not only direct identifiers such as names and contact information, but also any information that could identify individuals, including occupation, health status, and ideology, as PII. Unauthorized use of such information poses significant legal risks (European Parliament and Council of the European Union, 2016; Commission, 2021).

With the proliferation of Large Language Models (LLMs), the importance of PII management has further increased. While LLMs enable diverse applications such as translation, summarization, and question answering, they are typically pre-trained on massive corpora collected from the internet, including social media, which inevitably contain PII. Consequently, there is a risk that models may reproduce or leak PII (Huang et al., 2022; Lukas et al., 2023; Kim et al., 2023).

Various studies have been conducted on PII detection, and machine learning approaches are now predominantly used. However, while machine learning depends on the quantity and quality of training datasets, obtaining sufficient PII data for training is difficult due to legal constraints and ethical considerations.

Therefore, this study aims to develop a PII synthetic data generation method using LLMs and validate its effectiveness in improving PII detection accuracy through machine learning.

2 Related Work

2.1 PII Detection

Early research on PII detection primarily relied on regular expressions and rule-based methods (Sweeney, 1997; Beckwith et al., 2006). While effective for extracting explicit identifiers, these approaches have limitations in detecting context-dependent indirect personal information. Detection methods using Transformer-based models such as BERT and RoBERTa are becoming mainstream (Johnson et al., 2020; Elluri et al., 2021).

In specific domains, Mishra et al. have conducted a series of studies on PII detection for financial documents, leveraging synthetic data (Mishra et al., 2024, 2025). In the education domain, Shen et al. (2025) proposed a cost-efficient anonymization pipeline using a small model (GPT-4o-mini) for PII de-identification in educational data (Ji et al., 2025), demonstrating practical applicability.

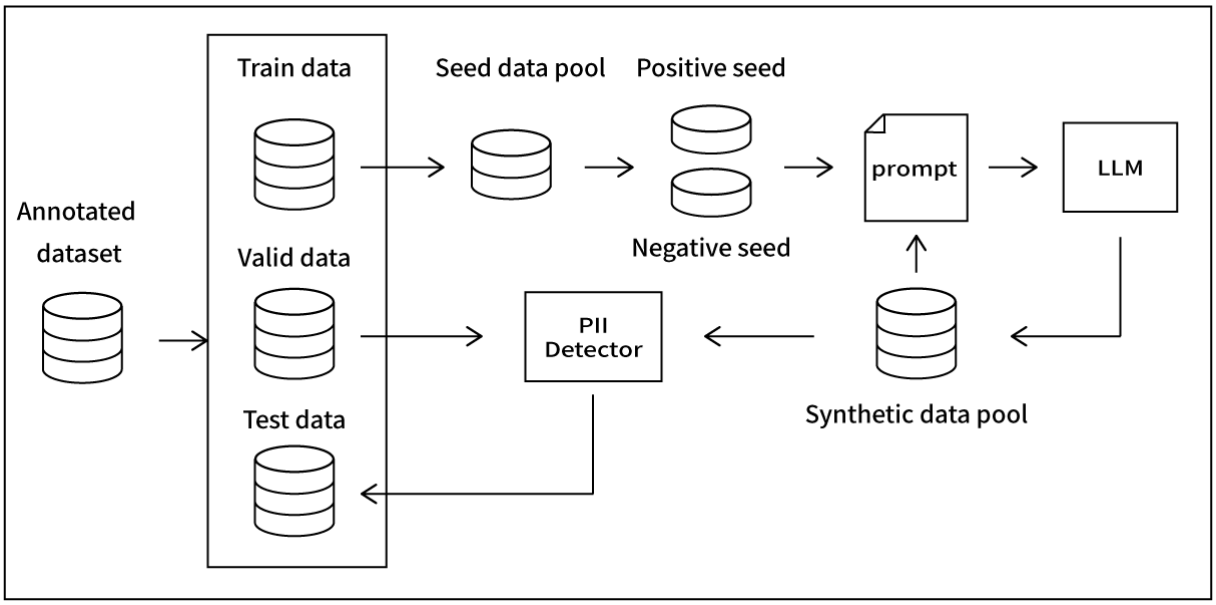

Figure 1: Overview of the proposed synthetic data generation framework for PII detection

2.2 PII Datasets

In PII detection research, datasets such as the Enron Email Dataset (Klimt and Yang, 2004) and The Text Anonymization Benchmark (TAB) (Pilán et al., 2022) have been widely used. However, these data focus on specific domains and may not adequately reflect the diversity and distribution of PII. In contrast, SPeDaC (Sensitive Personal Data Categories corpus) (Gambarelli et al., 2023) is a PII dataset labeled with a category system compliant with DPV (Data Privacy Vocabulary) based on real text from Wikipedia and Common Crawl, providing three different classification tasks (SPEDAC 1, 2, 3). Therefore, we use the SPeDaC dataset as a benchmark in this study.

2.3 Synthetic Data Generation

To overcome data scarcity, synthetic data generation using LLMs has gained attention. Santoso et al. (2024) improved low-resource NER performance by inducing open-source LLMs from few labeled examples to generate large amounts of artificial annotations (Santoso et al., 2024). Several studies have also been conducted on synthesizing datasets containing harmful information (Hui et al., 2024; Hartvigsen et al., 2022).

Research has also been conducted on PII synthesis in specific domains. Savkin et al. (2025) proposed synthetic PII datasets for medical and legal contexts (Savkin et al., 2025). Mishra et al. (2024) released a fully synthetic PII detection and anonymization dataset specialized for the financial domain, providing a foundation for safe model training and evaluation (Mishra et al., 2024). In the medical field, Šuvalov et al. (2025) developed a medical NER model effective even in low-resource language environments using synthetic healthcare data, achieving high accuracy while preserving patient privacy (Šuvalov et al., 2025).

This approach suggests applicability to general PII classification tasks. This study integrates these trends to develop a general PII synthetic data generation method using LLMs and validates its effectiveness through constructing PII detection models using the SPeDaC dataset.

3 Methods

This study develops a framework for generating diverse PII examples using LLMs, enabling high-accuracy PII classifier training even with small amounts of annotated data.

3.1 Prompt Design for Synthetic Data Generation

In this framework, we designed a few-shot-based prompt to generate diverse PII examples. Specifically, we randomly extract a specified number of text examples from an annotated dataset and save them as a seed data pool. Next, we create a prompt template. The information included in the prompt template is shown in Table 1.

Positive and negative examples are extracted from randomly selected text examples in the seed data pool. Text length is calculated from the aver-

Table 1: Prompt template components

Component	Description
Text length	Target length for generated text (calculated from average of positive examples)
Positive examples	Texts belonging to the specified class
Negative examples	Texts not belonging to the class (randomly extracted if multiple classes exist)
Previously generated	Previously generated texts (for preventing duplication)

Table 2: LLM generation parameters

Parameters	Value
temperature	1.5
top-p	0.9
top-k	60
max output tokens	800

Table 3: SPeDaC1 test dataset composition

Category	Samples	Percentage
Sensitive	1,087	49.0%
Non-sensitive	1,132	51.0%
Total	2,219	100.0%

age text length of extracted positive examples. If previously generated texts exist, they are included in the prompt to prevent the LLM from repeatedly generating the same text.

3.2 LLM

For synthetic data generation, we used gemini-2.5 series models. We selected the Gemini API because Google does not claim ownership of generated content, and the terms of service allow research use except for competing model development (Google, 2025). Due to the nature of generating synthetic data containing PII, we set all safety category thresholds (harassment, hate speech, sexual content, dangerous content) to `BLOCK_NONE` to avoid generation blocking by API-side safety filters.

4 Experimental Setup

In this study, we targeted SPeDaC1 and validated the effectiveness of LLM-based synthetic data generation. The SPeDaC1 dataset consists of 10,675 samples in total: Training set: 7,611, Validation set: 846, Test set: 2,218. We performed binary classification of texts containing PII (Positive) and texts not containing PII (Negative).

In this study, we randomly extracted an arbitrary number of samples from the Training set as a seed data pool. Subsequently, we extracted 10 positive examples and 10 negative examples, along with 10 randomly sampled examples from the most recent 100 generated samples. These were inserted into the prompt template and input to the LLM to generate 10,000 synthetic data samples: 5,000 Positive and 5,000 Negative. We used gemini-2.5-flash and gemini-2.5-flash-lite models as LLMs. The generation parameters are shown in Table 2.

We trained models using this synthetic data and evaluated them on the test dataset. The composition of the test dataset is shown in Table 3.

The classification model we used is RoBERTa (Robustly Optimized BERT Pre-training Approach) (Liu et al., 2019), which was also used in the SPeDaC paper. The hyperparameters used for training were set the same as in the paper. For evaluation, we adopted Accuracy as the primary metric.

5 Results

Figure 2 shows the detection performance as the number of training data samples varies across different conditions.

Table 4 shows the performance under each condition. Training on only 20–100 official data samples resulted in the model predicting all cases as Sensitive, with accuracy limited to 0.490. In contrast, synthetic data generated from the same amount of seeds (seed pool: 100, 9,000 samples) achieved an accuracy of 0.927, with a difference of only about 5 percentage points from official data with 10,000 samples (0.979). Even with seed pool: 10, accuracy of 0.873 was maintained, demonstrating that effective synthesis is possible from extremely small seeds.

6 Discussion

Training on 20–100 official data samples resulted in the model predicting all cases as Sensitive, with accuracy limited to 0.490. This indicates that appropriate decision boundaries cannot be learned from small amounts of real data alone. In contrast, 9,000 synthetic data samples generated from 100 seeds achieved an accuracy of 0.927, demonstrating that LLMs can generate diverse, high-quality PII examples. Even reducing the seed pool to 10 samples maintained an accuracy of 0.873, showing that practical synthesis is possible from extremely small seeds.

The advantages of our method lie in dramatic annotation cost reduction and a simple configuration consisting of few-shot prompts to high-

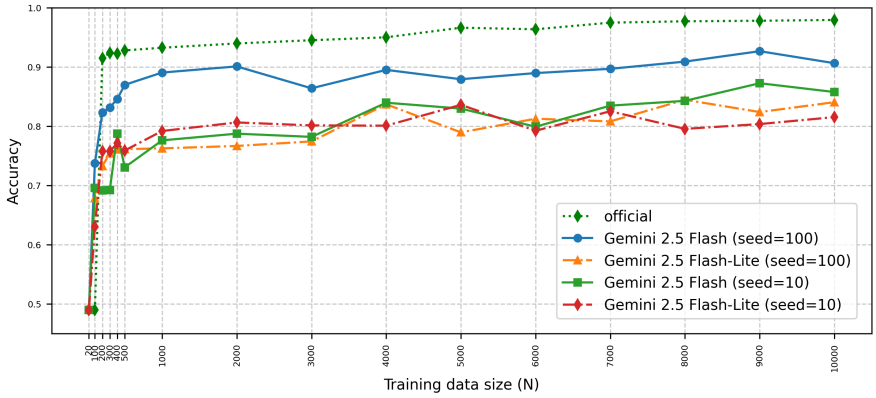

Figure 2: Synthetic data and performance with different numbers of seed data.

Table 4: Comparison of detection performance across training conditions

Data Type	Condition	Accuracy	Precision	Recall	F1-score	Support
Official data	20 samples	0.490	0.490	1.000	0.658	2,219
	100 samples	0.490	0.490	1.000	0.658	2,219
Synthetic data (gemini-2.5-flash)	Seed pool: 100	0.927	0.928	0.927	0.927	2,219
	Seed pool: 10	0.873	0.881	0.874	0.872	2,219

performance LLMs. Complex pipelines (data augmentation, back-translation, GANs, etc.) and domain-specific rule design are unnecessary, making implementation and maintenance easy and enabling rapid deployment to new domains. This flexibility is crucial for responding to rapidly changing privacy regulations.

Future challenges include extending to multi-class classification tasks (SPEDAC 2, 3), validation in other domains (medical, financial, etc.), and establishing quality evaluation metrics for synthetic data.

7 Conclusion

This study proposed a synthetic data generation framework using LLMs and demonstrated a method for generating practical PII detection datasets from extremely small seed data. In the SPeDaC1 benchmark, we generated 9,000 synthetic data samples from 100 seeds and achieved an accuracy of 0.927. This is within about 5 percentage points of official data with 10,000 samples (0.979), representing practical performance.

Training on only small amounts of real data (20–100 samples) was limited to an accuracy of 0.490, whereas synthetic data generated from the same amount of seeds achieved an accuracy of 0.927. This demonstrates that LLM-based synthetic data generation is an effective means of overcoming the limitations of small-data learning.

The main contributions of our method are four-fold: (1) achieving practical performance from extremely small seeds of 10–100 samples, dramatically reducing annotation costs; (2) demonstrating that the impact of seed pool size is limited (10 samples: 0.873 vs. 100 samples: 0.927); (3) clarifying the effectiveness of synthetic data through contrast with small real datasets; and (4) eliminating the need for complex pipelines and rule design through a simple configuration of few-shot prompts to LLMs.

These results enable rapid deployment of PII detection systems to new domains and low-resource languages. Future work includes extending to multi-class classification tasks (SPEDAC 2, 3), validation in other domains (medical, financial, etc.), and establishing quality evaluation metrics.

Limitations

This study has several limitations. First, the quality of synthetic data depends on LLM performance, requiring continuous evaluation. Second, we targeted only the binary classification task of SPeDaC1, and applicability to more complex multi-class classification (SPEDAC 2, 3) remains unverified. Third, the analysis of the performance gap between synthetic and real data (about 5 percentage points) is insufficient, and future investigation is needed regarding biases in generation quality by PII category and biases specific to synthetic data. Fourth, we used the English-based SPeDaC dataset, and generalizability to other languages remains unverified.

References

Bruce A. Beckwith, Rajeshwarri Mahaadevan, Ulysses J. Balis, and Frank Kuo. 2006. Development and evaluation of an open source software tool for deidentification of pathology reports. *BMC Medical Informatics and Decision Making*, 6(1):12.

European Commission. 2021. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS.

Lavanya Elluri, Sai Sree Laya Chukkapalli, Karuna Pande Joshi, Tim Finin, and Anupam Joshi. 2021. A BERT Based Approach to Measure Web Services Policies Compliance With GDPR. *IEEE Access*, 9:148004–148016.

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 (General Data Protection Regulation).

Gaia Gambarelli, Aldo Gangemi, and Rocco Tripodi. 2023. Is Your Model Sensitive? SPEDAC: A New Resource for the Automatic Classification of Sensitive Personal Data. *IEEE Access*, 11:10864–10880.

Google. 2025. Gemini API Additional Terms of Service. https://ai.google.dev/gemini-api/terms.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. 2022. Are Large Pre-Trained Language Models Leaking Your Personal Information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2038–2047, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, and Congrui Huang. 2024. ToxiCraft: A Novel Framework for Synthetic Generation of Harmful Information. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16632–16647, Miami, Florida, USA. Association for Computational Linguistics.

Zilyu Ji, Yuntian Shen, Kenneth R. Koedinger, and Jionghao Lin. 2025. Enhancing the De-identification of Personally Identifiable Information in Educational Data. *Journal of Educational Data Mining*, 17(2):55–85.

Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 214–221, New York, NY, USA. Association for Computing Machinery.

Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. 2023. ProPILE: Probing privacy leakage in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 20750–20762. Curran Associates, Inc.

Bryan Klimt and Yiming Yang. 2004. The Enron Corpus: A New Dataset for Email Classification Research. In *Machine Learning: ECML 2004*, pages 217–226, Berlin, Heidelberg. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. 2023. Analyzing Leakage of Personally Identifiable Information in Language Models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE Computer Society.

Kushagra Mishra, Harsh Pagare, Ranjeet Bidwe, and Sashikala Mishra. 2024. Synthetic Dataset for PII Detection and Anonymization in Financial Documents. 1.

Kushagra Mishra, Harsh Pagare, and Kanhaiya Sharma. 2025. A hybrid rule-based NLP and machine learning approach for PII detection and anonymization in financial documents. *Scientific Reports*, 15(1):22729.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Computational Linguistics*, 48(4):1053–1101.

Joan Santoso, Patrick Sutanto, Billy Cahyadi, and Esther Setiawan. 2024. Pushing the Limits of Low-Resource NER Using LLM Artificial Data Generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9652–9667, Bangkok, Thailand. Association for Computational Linguistics.

Maksim Savkin, Timur Ionov, and Vasily Konovalov. 2025. SPY: Enhancing Privacy with Synthetic PII Detection Dataset. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 236–246, Albuquerque, USA. Association for Computational Linguistics.

Hendrik Šuvalov, Mihkel Lepson, Veronika Kukk, Maria Malk, Neeme Ilves, Hele-Andra Kuulmets, and Raivo Kolde. 2025. Using Synthetic Health Care Data to Leverage Large Language Models for Named Entity Recognition: Development and Validation Study. *Journal of Medical Internet Research*, 27:e66279.

L. Sweeney. 1997. Guaranteeing anonymity when sharing medical data, the Datafly System. *Proceedings of the AMIA Annual Fall Symposium*, pages 51–55.

A Prompt Template for Synthetic Data Generation

Prompt Template for Sensitive Text Generation

Instruction: Generate a realistic text sample that contains personal or private information relevant to the evaluation scenario.

Target length: `{min_len}-{max_len}` characters (aim for ∼`{target_len}`).

Sensitive examples (with personal information).

[1] `{positive_example_1}`

... . . .

[10] `{positive_example_10}`

Non-sensitive examples (without personal information).

[1] `{negative_example_1}`

... . . .

[10] `{negative_example_10}`

Avoid generating text similar to the recent outputs.

[1] `{recent_generated_1}`

... . . .

[10] `{recent_generated_10}`

Generation requirements.

- Produce exactly one new sensitive text.
- Ensure the text differs from all provided examples while remaining realistic.
- Include explicit personal or private information consistent with the sensitive examples.

Figure 3: Prompt template used for synthetic data generation with LLM