LLM による合成データ生成を用いた個人情報分類の高精度化

江國翔太 [1,2,3]　曽傑 [4]　小橋洋平 [3]　小島武 [3]　岩澤有祐 [3]　松尾豊 [3]
[1] 株式会社 Vasculoid　[2]A-wave 株式会社　[3] 東京大学　[4] 成蹊大学
ekuni@vasculoid.com　jie-zeng@st.seikei.ac.jp
{yohei.kobashi, t.kojima, iwasawa, matsuo}@weblab.t.u-tokyo.ac.jp

概要

個人情報検出において，高精度な分類器の学習には大量のアノテーションデータが必要とされるが，プライバシー上の制約から実データの収集は困難である．本研究では，大規模言語モデル（LLM）を用いた合成データ生成によりこの課題を解決する手法を提案する．SPeDaC1 ベンチマークにおける実験では，わずか 100 件のシードデータから 9,000 件の合成データを生成し，Accuracy 0.927 を達成した．これは少量実データのみでの学習（0.490）と比較して大幅な改善であり，LLM による合成データ生成が少量データ学習の限界を克服する有効な手段であることを実証した．

1　はじめに

データ流通が世界規模で進む中，個人を特定し得る情報（Personally Identifiable Information, PII）の適切な管理は社会的信頼を支える重要な基盤である．

欧州委員会では氏名や連絡先といった直接的な識別子だけでなく，職業，健康状態，思想など個人を推定し得るあらゆる情報を PII として保護対象に含めており，これらの情報を無断で利用すれば重大な法的リスクが生じる [1, 2]．

大規模言語モデル（LLM）の普及によって，PII の管理の重要性は一層増した．LLM は翻訳・要約・質問応答など多彩な応用を可能にした一方で，SNS を含むインターネット上から収集した膨大なコーパスで事前学習されることが一般的であり，そこには PII が必然的に含まれている．その結果，モデルが PII を再現・漏洩する危険性が生じる [3, 4, 5]．

これまでに PII を検出する様々な研究が行われており，現在では機械学習的アプローチが主に用いられている．しかしながら，機械学習は学習に用いるデータセットの量と質に依存する一方で，法的制約・倫理的観点により，学習に十分な PII データを取得することは困難である．

そこで本研究では，LLM を活用した PII 合成データ生成手法を開発し，機械学習による PII 検出精度の向上に対する有効性を確認することを目的とする．

2　関連研究

2.1　PII 検出

個人情報検出（PII detection）の初期研究は，主に正規表現やルールベースの手法に依存していた [6, 7]．これらは明示的な識別子の抽出には有効である一方，文脈に依存した間接的な個人情報の検出には限界があり，Transformer アーキテクチャを基盤とした BERT や RoBERTa などのモデルを利用した検出手法が主流となりつつある [8, 9]．

特定のドメインに目を向けると，金融分野では，Mishra らが金融文書向けの PII 検出を対象に，合成データを活用した一連の研究を行っている [10, 11]．また，教育領域では Shen ら（2025）が教育データ中の PII 非特定化を対象に，小型モデル（GPT-4o-mini）を用いたコスト効率の高い匿名化パイプラインを提案しており [12]，実運用への適用性が示されている．

2.2　PII データセット

PII 検出研究では，Enron Email Dataset[13] や The Text Anonymization Benchmark (TAB)[14] などのデータセットが広く用いられてきた．しかし，これらのデータは特定のドメインにフォーカスしており，PII の多様性や分布を十分に反映していない可能性がある．一方，SPeDaC(Sensitive Personal Data Categories corpus)[15] は，Wikipedia と Common Crawl の実テキ

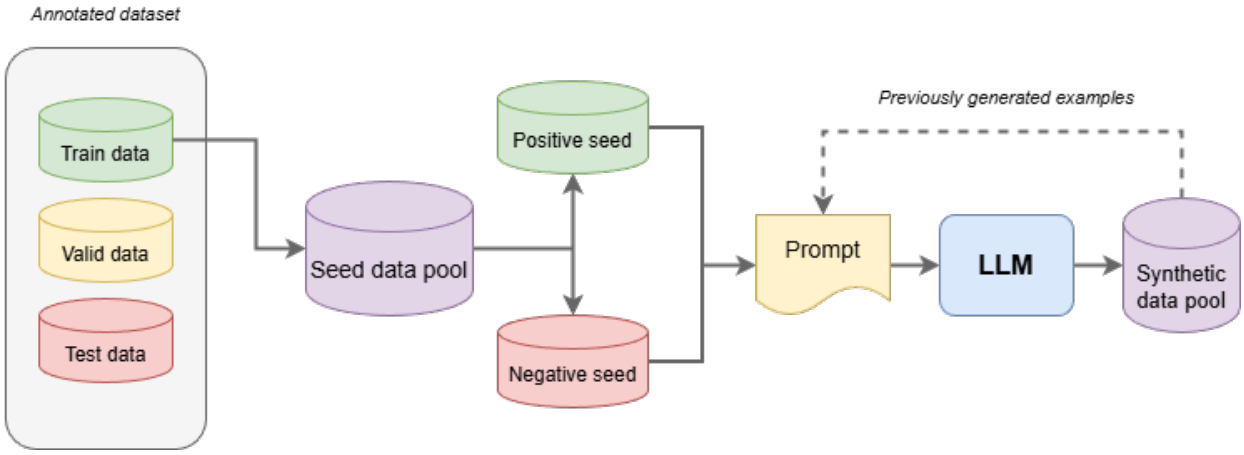

図 1　提案する PII 検出向け合成データ生成フレームワークの概要

ストをもとに DPV（Data Privacy Vocabulary）に準拠したカテゴリ体系でラベル付けされた PII データセットであり，異なる 3 つの分類タスク（SPEDAC 1, 2, 3）を提供している．そのため，本研究では，この SPeDaC のデータセットをベンチマークとして使用することとした．

2.3　合成データ生成

データ不足の克服に向けて，LLM を用いた合成データ生成が注目されている．Santoso ら（2024）は，少数ラベルの例からオープンソース LLM を誘導し，大量の人工アノテーションを生成して低リソース NER 性能を向上させた他 [16]，有害な情報を含むデータセットの合成に関しても，いくつかの研究が行われている [17, 18]．

特定ドメインの PII 合成の研究も行われており，Savkin ら（2025）は，医療と法務関連の合成 PII データセットを提案した [19]．Mishra ら（2024）は金融ドメイン特化の完全合成 PII 検出・匿名化データセットを公開し，安全にモデルを学習・評価できる基盤を提供した [10]．医療分野では Šuvalov ら（2025）が，合成ヘルスケアデータを用いて低リソース言語環境でも有効な医療 NER モデルを開発し，患者プライバシーを保ちながら高精度の実証を行っている [20]．

このアプローチは，汎用的な PII 分類タスクへの応用可能性を示唆する．本研究は，これらの流れを統合し，LLM による汎用的な PII 合成データ生成手法を開発し，SPeDaC データセットを用いた PII の検出モデルの構築によって，その有効性を検証する．

表 1　プロンプトテンプレート構成要素

要素	説明
テキスト長	生成するテキストの目標長（正例の平均から算出）
正例	指定したクラスに該当するテキスト群
負例	クラスに該当しないテキスト群（複数クラスある場合はランダムに抽出）
生成済例	既に生成されたテキスト群（重複防止用）

3　提案手法

本研究では，LLM を用いて多様な PII 例を生成するフレームワークを開発し，少量のアノテーションデータでも高精度な PII 分類器学習を実現する．

3.1　合成データ生成のためのプロンプト設計

本フレームワークでは，PII の多様な例を生成するための Few-shot ベースのプロンプトを設計した．具体的には，アノテーション済みデータセットからランダムに指定した件数のテキスト例を抽出し，それらをシードデータプールとして保存する．次に，プロンプトテンプレートを作成した．プロンプトテンプレートに含まれる情報を表 1 に示す．

正例と負例はシードデータプールからランダムに指定した件数のテキスト例から抽出する．テキスト長は，抽出した正例の平均テキスト長から算出する．また，既に生成されたテキスト群が存在する場合はそれらをプロンプトに含めることで，LLM が同じテキストを繰り返し生成することを防止する．

3.2　LLM

合成データ生成には，gemini-2.5 系モデルを使用した．Gemini API を選択した理由は，生成されたコンテンツの所有権を Google が主張せず，利用規約

上で競合モデル開発以外の研究目的での使用が可能であったためである [21]．なお，PII を含む合成データを生成する性質上，API 側のセーフティフィルタによる生成ブロックを回避するため，全てのセーフティカテゴリ（ハラスメント，ヘイトスピーチ，性的コンテンツ，危険コンテンツ）の閾値を BLOCK_NONE に設定した．

4 実験

本研究では，SPeDaC1 を対象とし，LLM による合成データ生成の有効性を検証した．SPeDaC1 のデータセットは，Training set: 7,611, validation set: 846, test set: 2,218 の計 10,675 件で構成されており，PII を含むテキスト（Positive）と PII を含まないテキスト（Negative）の二値分類を行った．

4.1 実験設定

本研究では，Training set から任意の件数を seed data pool としてランダムに抽出した．その後，正例:10, 負例:10 に加えて，直近の生成済みの 100 例から 10 件をランダムに抽出した．これらをプロンプトテンプレートに挿入して LLM に入力し，Positive: 5,000 件，Negative: 5,000 件の合計 10,000 件の合成データを生成した．LLM には，gemini-2.5-flash および gemini-2.5-flash-lite モデルを使用した．生成時のパラメータは表 2 に示す．この合成データを用いてモデルの学習を行い，テストデータセットで評価を行った．使用した分類モデルは，SPeDaC の論文でも使用されている RoBERTa (Robustly Optimized BERT Pre-training Approach)[22] である．学習に使用したハイパーパラメータは論文と同様に設定した．評価には，Accuracy を主要指標として採用した．また，合成データの品質を評価するために，語彙多様性（Unique 2-gram ratio），分布類似性（Jensen-Shannon divergence），内部多様性（Self-BLEU）を計算した．

表 2 LLM 生成パラメータ

パラメータ	値
temperature	1.5
top-p	0.9
top-k	60
max output tokens	800

4.2 結果

図 2 に各条件で学習させるデータの件数を変化させた検出性能を示す．表 3 に各条件の性能を示す．表 4 に，合成データの品質を語彙多様性，分布類似性，内部多様性の観点から分析した結果を示す．公式データ 20〜100 件のみでの学習では，モデルが全事例を Sensitive と予測し，Accuracy 0.490 に留まった．一方，同量のシードから生成した合成データ（seed pool: 100，9,000 件）では Accuracy 0.927 を達成し，公式データ 10,000 件（0.979）との差は約 5% ポイントに抑えられた．seed pool: 10 でも Accuracy 0.873 を維持し，極小シードからの有効な合成が可

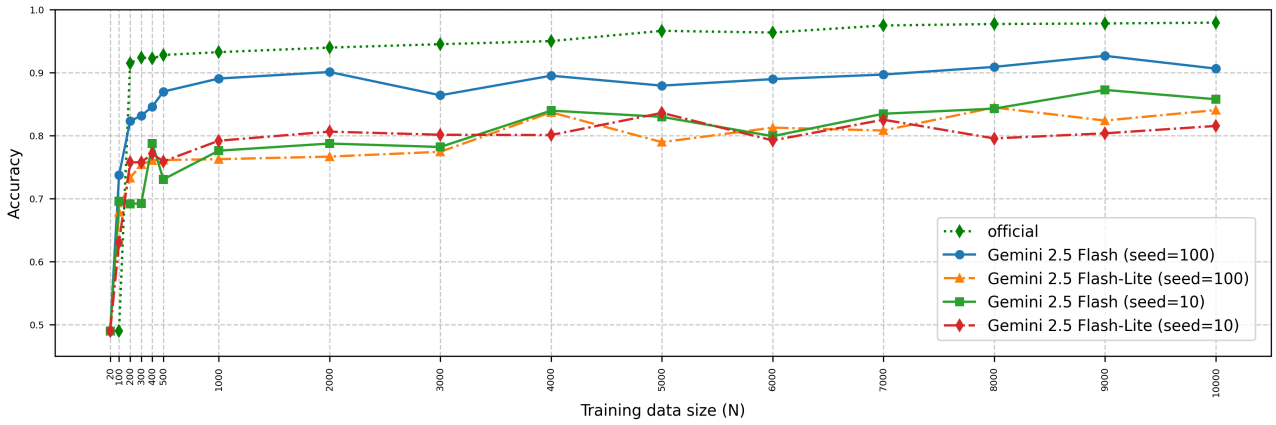

図 2 シードデータ件数別の合成データと検出性能

表 3 各学習条件における検出性能の比較

データ種別	条件	Accuracy	Precision	Recall	F1-score	Support
公式データ	20 件	0.490	0.490	1.000	0.658	2,219
	100 件	0.490	0.490	1.000	0.658	2,219
合成データ (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

表 4 合成データの品質分析

データセット	Unique 2-gram	JS Div.	Self-BLEU
実データ (Train)	0.499	-	0.314
合成 (seed=10, Flash)	0.360	0.554	0.326
合成 (seed=10, Lite)	0.317	0.531	0.419
合成 (seed=100, Flash)	0.396	0.527	0.323
合成 (seed=100, Lite)	0.287	0.538	0.445

能である．

5 考察

公式データ 20〜100 件での学習では，モデルが全事例を Sensitive と予測し，Accuracy 0.490 に留まった．これは少量実データのみでは適切な決定境界を学習できないことを示す．対照的に，100 件のシードから生成した 9,000 件の合成データでは Accuracy 0.927 を達成し，LLM が多様で質の高い PII 例を生成できることを実証した．seed pool を 10 件に削減しても Accuracy 0.873 を維持でき，極小シードからの実用的な合成が可能である．

語彙多様性（Unique 2-gram ratio）は，合成データが実データの 72–80%程度であり，Flash が Flash-Lite を一貫して上回った．JS divergence は全条件で 0.5 以上と高く，合成データと実データの分布には差異が存在する．一方，Self-BLEU（低いほど多様）では Flash（0.32–0.33）が Flash-Lite（0.42–0.45）を大きく上回り，Flash の方が繰り返しの少ない多様なテキストを生成することが確認された．

これらの結果から，合成データは実データの分布を完全には再現していないものの，PII 検出タスクにおいて有効な学習データとして機能することが示された．

本手法の利点は，大規模アノテーションコスト削減と，高性能 LLM への Few-shot プロンプトというシンプルな構成にある．従来の複雑なパイプライン（データ拡張，バックトランスレーション，GAN など）やドメイン固有のルール設計が不要であり，実装・保守が容易で，新規ドメインへ迅速に展開できる．この柔軟性は，プライバシー規制の急速な変化への対応において重要である．

6 おわりに

本研究では，LLM を用いた合成データ生成フレームワークを提案し，極小シードデータから実用的な PII 検出用データセットを生成する手法を実証した．SPeDaC1 ベンチマークにおいて，100 件のシードから 9,000 件の合成データを生成し，Accuracy 0.927 を達成した．これは公式データ 10,000 件（0.979）と約 5%ポイント差に抑えられ，実用的な性能である．少量実データ（20〜100 件）のみでの学習が Accuracy 0.490 に留まったのに対し，同量のシードから生成した合成データでは Accuracy 0.927 を達成した．これは，LLM による合成データ生成が少量データ学習の限界を克服する有効な手段であることを示す．

本手法の主要な貢献は 4 点である：(1) 10〜100 件の極小シードから実用的性能を実現し，アノテーションコストを劇的に削減，(2) seed pool サイズの影響が限定的（10 件: 0.873 vs 100 件: 0.927）であることを実証，(3) 少量実データとの対比により合成データの有効性を明確化，(4) LLM への Few-shot プロンプトというシンプルな構成により，複雑なパイプラインやルール設計を不要化．

これらの成果は，新規ドメインや低リソース言語への迅速な PII 検出システム展開を可能にする．今後は，多クラス分類タスク（SPEDAC 2, 3）への拡張，他ドメイン（医療，金融など）での検証，品質評価指標の確立を進める．

制限事項

本研究にはいくつかの限界がある．第一に，合成データの品質は LLM の性能に依存するため，他の商用モデルやオープンモデルを用いた継続的な評価が必要である．第二に，SPeDaC1 の二値分類のみを対象としており，多クラス分類（SPEDAC 2, 3）や他の PII ベンチマークへの適用性を今後検討していく必要がある．また，PII カテゴリ別の生成品質の偏りや合成データ特有のバイアスについても調査が必要である．

参考文献

[1] European Parliament and Council of the European Union. 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), 2016.

[2] European Commission. 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, 2021.

[3] Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are Large Pre-Trained Language Models Leaking Your Personal Information? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 2038–2047, Abu Dhabi, United Arab Emirates, February 2022. Association for Computational Linguistics.

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

[5] Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon, and Seong Joon Oh. ProPILE: Probing privacy leakage in large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, **Advances in Neural Information Processing Systems**, Vol. 36, pp. 20750–20762. Curran Associates, Inc., 2023.

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

[7] Bruce A. Beckwith, Rajeshwarri Mahaadevan, Ulysses J. Balis, and Frank Kuo. Development and evaluation of an open source software tool for deidentification of pathology reports. **BMC Medical Informatics and Decision Making**, Vol. 6, No. 1, p. 12, March 2006.

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

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

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

[11] Kushagra Mishra, Harsh Pagare, and Kanhaiya Sharma. A hybrid rule-based NLP and machine learning approach for PII detection and anonymization in financial documents. **Scientific Reports**, Vol. 15, No. 1, p. 22729, July 2025.

[12] Zilyu Ji, Yuntian Shen, Kenneth R. Koedinger, and Jionghao Lin. Enhancing the De-identification of Personally Identifiable Information in Educational Data. **Journal of Educational Data Mining**, Vol. 17, No. 2, pp. 55–85, September 2025.

[13] Bryan Klimt and Yiming Yang. The Enron Corpus: A New Dataset for Email Classification Research. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, **Machine Learning: ECML 2004**, pp. 217–226, Berlin, Heidelberg, 2004. Springer.

[14] Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. **Computational Linguistics**, Vol. 48, No. 4, pp. 1053–1101, February 2022.

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

[16] Joan Santoso, Patrick Sutanto, Billy Cahyadi, and Esther Setiawan. Pushing the Limits of Low-Resource NER Using LLM Artificial Data Generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Findings of the Association for Computational Linguistics: ACL 2024**, pp. 9652–9667, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[17] Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, and Congrui Huang. ToxiCraft: A Novel Framework for Synthetic Generation of Harmful Information. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 16632–16647, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

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

[19] Maksim Savkin, Timur Ionov, and Vasily Konovalov. SPY: Enhancing Privacy with Synthetic PII Detection Dataset. In Abteen Ebrahimi, Samar Haider, Emmy Liu, Sammar Haider, Maria Leonor Pacheco, and Shira Wein, editors, **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)**, pp. 236–246, Albuquerque, USA, April 2025. Association for Computational Linguistics.

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

[21] Google. Gemini API Additional Terms of Service. https://ai.google.dev/gemini-api/terms, September 2025.

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

A 合成データ生成用プロンプトテンプレート

合成データ生成に使用したプロンプトテンプレートを以下に示す.

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).

- `{positive_example_1}`
- ...
- `{positive_example_10}`

Non-sensitive examples (without personal information).

- `{negative_example_1}`
- ...
- `{negative_example_10}`

Avoid generating text similar to the recent outputs.

- `{recent_generated_1}`
- ...
- `{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.