大規模言語モデルによる臨床テキストの非識別化

矢田 竣太郎 [1] 小林 和馬 [2,3] 伊藤 沙紀子 [4] 小田 悠介 [3] 相澤 彰子 [3]
[1] 筑波大学 [2] 国立がん研究センター
[3] 国立情報学研究所 大規模言語モデル研究開発センター [4] 東京科学大学
yada@slis.tsukuba.ac.jp kazumkob@ncc.go.jp
itoh.sakiko@tmd.ac.jp yusuke.oda@naist.ac.jp aizawa@nii.ac.jp

概要

医療分野における大規模言語モデル（LLM）の安全な利活用には、学習および推論段階におけるプライバシー保護技術の確立が急務である。特に、診療テキストに含まれる氏名等の個人識別情報の検出と除去（非識別化）は重要な課題である。本研究では、擬似的な個人識別情報を付与した約 28 万件の日本語診療テキストからなる大規模なデータセットを構築した。このデータセットを用いて、130 億パラメータの日本語医療特化 LLM に対して教師ありファインチューニングを行い、非識別化に特化したモデルを開発した。結果、氏名等の個人特定性の高い記述（識別子）に対して F1 値 98.5%を達成するとともに、LLM ベースの非識別化では文脈に応じた柔軟な処理が可能であることが示唆された。

1 はじめに

大規模言語モデル（LLM）の医療分野への応用が急速に進むなかで、診療テキストなどの機微情報を安全に利活用するためのプライバシー保護技術の確立が急務である。一般的に、診療テキストに含まれる医療情報は、患者が自身の傷病の快復等を目的として医療機関を受診し、診療を受けるために提供したものである。そのため、LLM の学習や推論段階において、診療テキストに含まれる個人識別情報が外部に漏洩し、これが不適切に取り扱われることによって、患者本人に対する差別や偏見といった深刻な権利利益の侵害が生じる恐れがある。

こうしたプライバシー保護技術の中核となるのが**非識別化（De-identification）**である。非識別化とは、テキストから**個人識別情報（Personally Identifiable Information; PII）**を検出し、これを削除または置換することで、当該テキスト単体からも、あるいは当該テキストと他の情報との組み合わせによっても、特定の個人が識別されないように加工する技術的安全管理措置である [1]。

対象となる PII は、個人識別性の程度と潜在的な権利利益侵害の態様に応じて、複数の類型に分類される。まず、氏名など、記述単体で特定の個人の識別に至ることができるものとしての**識別子**である。次に、年齢、生年月日、住所といった、他の記述との組み合わせにより特定の個人の識別に至ることができる**準識別子**である。さらに、電話番号といった**連絡先情報**のように、漏洩することで個人に直接被害が及ぶものや、他の情報との連結に用いられる**連結符号**、および個人情報保護法で定められる**個人識別符号**といったものが含まれる。

重要な点として、PII が有する個人識別性は文脈や母集団に依存した相対的な概念であり、特定の記述の PII への該当性を一律のルールベースで判定することは容易ではない。そのため、LLM などを活用した非識別化技術の確立が急務であった。実際、非識別化は医療言語処理の主流タスクで、主に固有表現抽出の枠組みが採用されてきた [2, 3]。LLM を用いた手法も提案されているが、英語テキストでのプロンプトエンジニアリング適用事例が多い [4]。

そこで本研究では、まず擬似的な PII を付与した約 28 万件の日本語診療テキストからなる大規模なデータセットを構築した。これは、日本語での医療分野のプライバシー保護技術の研究開発に資する初の大規模なデータセットである。さらに、本データセットを用いて 130 億パラメータの日本語医療特化 LLM（`SIP-jmed-llm-3-13b-32k-base`）に対して教師ありファインチューニングを行い、非識別化に特化したモデルを開発した。その結果、部分一致を許容すれば、氏名等の個人特定性の高い記述（識別子）に対して F1 値 98.5%を達成するとともに、LLM ベースの非識別化では文脈に応じた柔軟な処理が可能であることが示唆された。

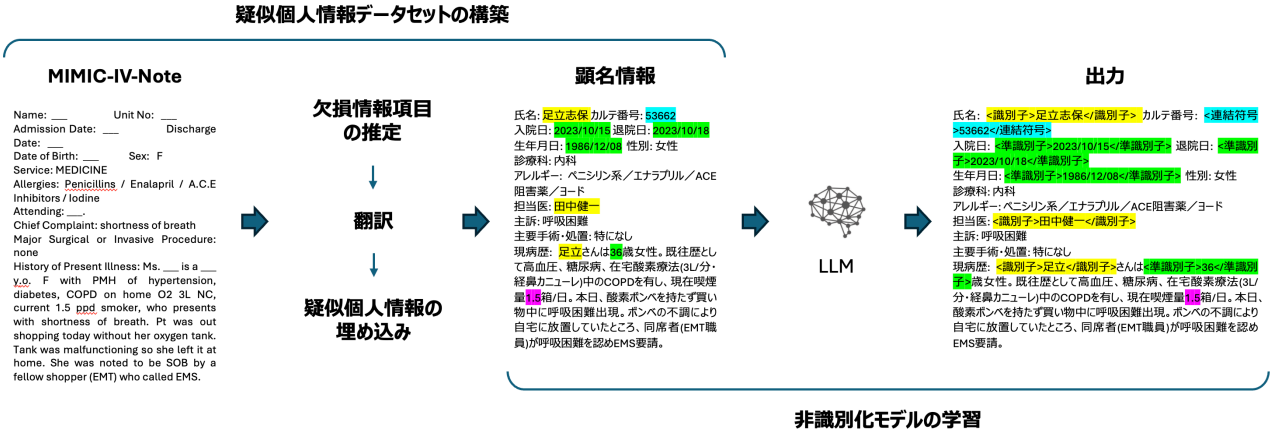

図 1　本研究の流れ図。英語の MIMIC-IV-Notes を日本語に翻訳し、文脈に合わせた擬似個人情報を埋め込む。この擬似的な顕名診療録から、個人識別情報を類型ごとに抽出できる非識別化モデルを構築した。

2　手法

2.1　擬似個人情報データセットの構築

はじめに、英語の診療テキストデータセットである MIMIC-IV-Note [5] を取得した。MIMIC-IV-Note は約 33 万件の救急部門の退院サマリーを含み、非識別化が施された後、氏名等の PII が _____ のようにマスクされた状態で条件付き公開されている。これを擬似的な PII が付与された日本語診療テキストデータセットに変換するため、以下の工程を実施した（**図 1** 参照）。

まず、LLM を用いて MIMIC-IV-Note のオリジナルテキストを分析し、どの箇所に氏名、住所、生年月日、検査日、病院名といった情報項目がマスクされたかを前後の文脈から推定した。推定結果に基づき、患者氏名の該当箇所には `{{Patient Name}}` といったプレースホルダを埋め込んだ。次に、プレースホルダ付きの英語診療テキストを日本語に翻訳した。その後、予め作成した日本人の氏名等からなる擬似個人情報リスト（詳細は**付録 A** を参照）から、翻訳後のテキストに含まれるプレースホルダを、性別や年齢層といった属性情報に基づいて擬似的な日本人の PII に置換した（例：`{{Patient Name}}` → `{{則本詩織}}`）。擬似個人情報の対応付けは MIMIC-IV-Note の患者単位（`subject_id`）で一貫性を保つよう行った。

MIMIC-IV-Note のオリジナルテキストを検討する過程で、検査値など必ずしも PII に該当しない記述もマスクされていることが判明した。こうした情報については、LLM を用いて前後の文脈に基づき自然な記述に復元した。以上の処理は、LLM として DeepSeek-V3 [6]（v3-0324）を用いて実施した。

2.2　PII 類型の定義と分類

処理対象としての PII について、個人識別性の高低と潜在的な権利利益侵害の態様に応じて、以下の 5 つの類型を定義した。

識別子　氏名など、単体で特定個人を識別できる記述（例: 患者氏名、患者家族の氏名、医師氏名等）。

準識別子　組み合わせにより個人識別可能となる記述（例: 姓のみ、名のみ、イニシャル、郵便番号、住所、生年月日、性別、所属、学歴、職歴、入退院日、受診日、検査日等の日付、医療機関名、クリニック名等）。

個人識別符号　旅券番号、基礎年金番号、免許証番号、住民票コード、マイナンバー、各種保険証番号といった公的機関が割り振る番号。

連結符号　カルテ番号等、連結用 ID として利用される記述。

連絡先情報　電話番号等の本人連絡先。

続いて、前工程で復元された個々の記述について、元のプレースホルダの情報を参照しながら、それぞれがどの PII 類型に該当するかを Qwen3-32B [7] で推定した。推定結果については人手で確認を行い、確認可能な範囲で修正を行った。最終的に、PII 類型および出現位置がメタ情報として付与された擬似個人情報を含む、合計 281,993 件の日本語診療テキストからなるデータセットを構築した。以降、擬似個人情報がテキストに出現する状態を「顕名」と表現する。

表 1　データセットにおける各ラベルの出現頻度（学習データの頻度は 5 万件のランダムサンプリングに基づく換算値）

PII 類型	学習用（頻度）	テスト用
識別子	97,617	9,586
連結符号	61,230	5,956
準識別子	1,675,526	147,424
連絡先情報	24,329	2,272
個人識別符号	162	12

2.3　非識別化用データセットの構築

本研究では、非識別化を次のような固有表現抽出タスクとして定義する。顕名の診療テキストを入力として、これに含まれる PII を検出し、適切な PII 類型を判定して、タグで囲んだテキストを出力する。ただし、PII に該当しない記述や文章については、入力された診療テキストから改変されないものとする。この定義に基づき、正解データとして、各診療テキストに含まれる個々の PII をそれぞれ該当する PII 類型を表現するタグ（例：<識別子>則本詩織</識別子>）で囲んだテキストを作成した（**図 1**）。

全体 281,993 件の診療テキストのうち、検証用およびテスト用としてそれぞれ 5,000 件を割り当てた。ただし、本データセットには一人の患者につき複数の診療テキストが含まれるため、検証用およびテスト用は単一の診療テキストのみを有する患者のみを対象として、患者単位でのデータリークが生じないよう配慮した。残りの 271,993 件のうち 50,000 件をランダムに抽出し、学習用データセットとした。5 つの PII 類型の分布は**表 1** に示す通りである。

2.4　非識別化モデルの学習

日本語医療分野に特化した 130 億パラメータのモデル（`SIP-jmed-llm-3-13b-32k-base`）[1] をベースモデルとして、学習用データセットを用いた教師ありファインチューニング（Supervised Fine-Tuning; SFT）を実施した（**図 1**）。このベースモデルは、`LLM-jp-3.1-13b`[2] を日本語医療ドメインに特化するよう継続事前学習したものであり、32k トークンのコンテクスト長を有している。学習では Cross Entropy Loss を用いて、入力された診療テキストから PII をタグで囲んだテキストを出力するようにモデルを訓練した。具体的なハイパーパラメータ等の設定は**付録 C** に示す。

2.4.1　評価指標

モデルの性能評価にあたり、固有表現抽出タスクで標準的に用いられる精度、再現率、F1 値を、タグ付けられた PII 記述ごと（Entity-level）に算出する。加えて、医療文書の非識別化実務を自動化するにあたり、文書単位（Record-level）での処理完全性が重要視されることから、次のような 3 つの安全性評価指標を導入する。

完全抽出率:　レコードに含まれる当該 PII 類型の個人識別情報を、一つ残らず全て検出できた（文書単位再現率＝ 1）レコードの割合。

無誤抽出率:　レコードに対してモデルが当該 PII 類型として抽出した箇所が、全て正解であったレコード（文書単位精度＝ 1）の割合。

完答率:　検出漏れ（False Negative）も過検出（False Positive）もなく、完全に正解と一致したレコード（文書単位 F1 値＝ 1）の割合。

さらに、抽出されたタグの位置と PII 類型に関するラベルの整合性に基づいて、以下の 3 つの評価基準を定義した。

1. **Strict**：位置とラベルが完全に一致した場合のみ正解とする。
2. **Relaxed**：ラベルが一致し、かつ位置が 1 文字以上重複していれば正解とする。
3. **Label-Relaxed**：ラベルの種類に関わらず、位置が部分的に重複していれば個人情報の検出に成功したとみなし、正解とする。

これら評価基準の違いを**表 2** にまとめた。本稿では主に Label-Relaxed 基準の結果を中心に報告し、Strict および Relaxed 基準の結果は付録 D に示す。

表 2　評価基準の定義と許容範囲

評価基準	スパン（位置）	ラベル（分類）
Strict	完全一致	完全一致
Relaxed	部分一致	完全一致
Label-Relaxed	部分一致	不問

3　結果と考察

SFT を実施した非識別化モデル（2.4 節）を評価した。前項で述べた各評価値（2.4.1）の報告に加え、定性的に非識別化事例も紹介する。

1）https://sites.google.com/nii.ac.jp/sip3e-2/公開リソース 2

2）https://llmc.nii.ac.jp/topics/llm-jp-3-1_instruct4/

図 2　Label-Relaxed 評価における混同行列

表 3　PII 類型別評価結果（Label-Relaxed 基準）

PII 類型	精度	再現率	F1 値
識別子	98.7%	98.2%	**98.5%**
準識別子	96.5%	94.8%	95.7%
連結符号	97.2%	98.0%	97.6%
連絡先情報	96.3%	96.0%	96.2%
個人識別符号	90.9%	100.0%	95.2%

表 4　文書単位の安全性評価（Label-Relaxed 基準）

PII 類型	完全抽出率	無誤抽出率	完答率
識別子	97.68%	98.58%	97.00%
準識別子	42.64%	61.38%	30.26%
連結符号	98.54%	98.58%	97.26%
連絡先情報	94.82%	94.76%	91.32%
個人識別符号	100.00%	90.91%	92.31%

3.1　定量評価

テスト用データセット 5,000 件に対する評価結果を表 3 に示す。Label-Relaxed 基準において、タスク上もっとも重要な「識別子」（氏名等）の F1 値は 98.5%、「連結符号」（カルテ番号等）は 97.6%と高い性能を示した。それ以外のラベルに対しても、一貫して 95%以上の高い F1 値が得られた。

また、レコード単位の統計では、Label-Relaxed 基準における「識別子」の完全抽出率は 97.68%であった。これは、ラベル分類の厳密さを問わなければ、約 98%の文書において、含まれる氏名等の識別子を完全に検出・除去可能であることに相当し、実務的な非識別化支援ツールとしての有用性を示唆する。

図 2 にラベル推定結果の混合行列を示す。正解データにおける「識別子」が、モデルによって「準識別子」と予測された事例が 1,133 件確認された。この誤りは「識別子」において大部分を占めており、モデルは当該スパンが保護すべき機微情報であることは正しく検知できているものの、それが特定の個人を直接識別する情報（識別子）か、組み合わせによって識別し得る情報（準識別子）かの判定において、文脈上の曖昧さに影響を受けていることが示唆される。一方で「準識別子」に関しては、検出漏れが 7,624 件、過検出が 5,083 件と比較的多く、この類型それ自体の曖昧性が高いものと思われる。

3.2　定性事例分析

非識別化モデルによる PII 抽出結果を目視で確認していたところ、「準識別子」として下記のように比較的長いスパンで抽出される事例が散見された。

> <準識別子>大学教授として勤務中。摂食障害専門の臨床心理士資格保有。週 3 回の軽いウォーキングを習慣としている。</準識別子>

こういった記述は、暗黙ながらも個人識別性の高いものといえる。単語や句にとどまらない複雑な表現を文脈に応じて柔軟に検出できている点では、LLM をベースとした本手法の可能性が見込まれる。

4　おわりに

本研究では、医療分野における生成 AI の利活用に係るプライバシー保護技術の研究開発を促進するために、MIMIC-IV-Note を基盤とした大規模な日本語擬似個人情報データセットを構築した。さらに、本データセットを活用することで、高精度で個人識別情報を非識別化できる LLM を構築した。評価の結果、特に氏名等の識別子に対して F1 値 98.5%を達成し、自動化を期待できる性能を確認した。今後の課題として、準識別子の定義の精緻化および、より多様な臨床文書への汎化性能の検証が挙げられる。

謝辞

本研究は厚生労働科学研究費 25IA1012、および内閣府・戦略的イノベーション創造プログラム（SIP）「統合型ヘルスケアシステムの構築における生成 AI の活用」の支援を受けて実施した。

参考文献

[1] Raphaël Chevrier, Vasiliki Foufi, Christophe Gaudet-Blavignac, Arnaud Robert, and Christian Lovis. Use and understanding of anonymization and DE-identification in the biomedical literature: Scoping review. **Journal of Medical Internet Research**, Vol. 21, No. 5, p. e13484, 31 May 2019.

[2] Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. Automatic de-identification of textual documents in the electronic health record: a review of recent research. **BMC Medical Research Methodology**, Vol. 10, No. 1, p. 70, 2 August 2010.

[3] Amber Stubbs and Özlem Uzuner. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. **Journal of Biomedical Informatics**, Vol. 58 Suppl, pp. S20–S29, 1 December 2015.

[4] Bayan Altalla', Sameera Abdalla, Ahmad Altamimi, Layla Bitar, Amal Al Omari, Ramiz Kardan, and Iyad Sultan. Evaluating GPT models for clinical note de-identification. **Scientific Reports**, Vol. 15, No. 1, p. 3852, 31 January 2025.

[5] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV, 2024.

[6] DeepSeek-AI. DeepSeek-V3 technical report. **arXiv [cs.CL]**, 26 December 2024.

[7] An et al. Yang. Qwen3 technical report. **arXiv [cs.CL]**, 14 May 2025.

A　擬似個人情報リストの作成

データセットの再マスクに用いた擬似個人情報リストは次のように作成した。まず、実在する人名データ[3]を元にした日本人の苗字と名前、性別の組み合わせを大量に生成（10 万人に数人単位の同姓同名を含む）した。次に、擬似個人情報作成ツール[4]で電話番号、住所などを生成した。

顕名化処理においては、MIMIC-IV-Note から患者ごとの性別、年齢、誕生日を LLM により別途推定し、この推定された患者プロフィールにマッチする擬似個人情報をリストから紐づけるようにした。

B　プロンプト例

本研究の評価実験で使用したモデルへの入力プロンプトと出力の例を以下に示す。

System Prompt　以下の診療テキストから個人識別情報を抽出してください。

User Input　氏名: 足立志保 \n カルテ番号: 53662 \n ... （以下、診療テキスト）

Model Output　氏名:<識別子>足立志保</識別子> \n カルテ番号: <連結符号>53662</連結符号> \n ...

C　モデル学習設定

SFT における主な学習設定は 5 の通りである。なお、PII 類型タグ（<識別子> など）は通常のテキストトークンとして処理した。

表 5　学習および生成のハイパーパラメータ

項目	設定値
Base Model	LLM-jp-3.1-13b-32k-base
精度	bfloat16
Max Tokens	32,768
Temperature	0.0
Top-p	1.0
Stop Tokens	<EODLLM-jp>|, </s>
Tensor Parallel Size	8

D　詳細な評価結果

Strict 基準および Relaxed 基準の評価結果を表 6〜表 9 に示す。

表 6　PII 類型別評価結果（Strict 基準）

PII	精度	再現率	F1 値
個人識別符号	90.9%	83.3%	87.0%
準識別子	91.9%	90.7%	91.3%
識別子	90.0%	84.7%	87.2%
連結符号	95.4%	94.9%	95.2%
連絡先情報	87.9%	86.4%	87.2%

表 7　PII 類型別評価結果（Relaxed 基準）

PII 類型	精度	再現率	F1 値
個人識別符号	90.9%	83.3%	87.0%
準識別子	95.6%	94.4%	95.0%
識別子	91.7%	86.3%	88.9%
連結符号	96.7%	96.2%	96.5%
連絡先情報	94.9%	93.3%	94.1%

表 8　文書単位の安全性評価（Strict 基準）

PII 類型	完全抽出率	無誤抽出率	完答率
識別子	80.98%	88.58%	74.23%
準識別子	32.82%	42.24%	21.30%
連結符号	96.52%	97.50%	95.28%
連絡先情報	84.59%	86.70%	80.76%
個人識別符号	83.33%	90.91%	76.92%

表 9　文書単位の安全性評価（Relaxed 基準）

PII 類型	完全抽出率	無誤抽出率	完答率
識別子	82.42%	90.10%	75.40%
準識別子	40.16%	52.50%	24.98%
連結符号	97.52%	98.38%	96.10%
連絡先情報	91.10%	93.08%	86.37%
個人識別符号	83.33%	90.91%	76.92%

3）https://nextvitz.com/jp/name-random-1.php
4）https://hogehoge.tk/personal/