大規模言語モデルの事前学習用コーパスにおける要配慮個人情報の検出

源怜維 [1]　小田悠介 [2]　河原大輔 [1,2]
[1] 早稲田大学理工学術院　[2] 国立情報学研究所 大規模言語モデル研究開発センター
{ray@akane.,dkw@}waseda.jp　odashi@nii.ac.jp

概要

法的に取得が制限される要配慮個人情報は，大規模言語モデル（LLM）の構築に必要な大規模な事前学習用コーパスに含まれる可能性があり，その検出とフィルタリングは重要な課題である．本研究では，文章内の要配慮個人情報を検出するために，要配慮個人情報データセットを構築し，機械学習モデルを学習させた．その結果，要配慮個人情報のジャンルと関係する情報を高速に検出する判定器を構築できた．

1　はじめに

LLM の構築には大規模な事前学習用コーパスが必要であり，その収集手段として Web クローリングが広く利用されている．通常，クローラは自由に取得可能なデータを網羅的に探索するが，この過程で実際には利用が制限される情報を誤って収集してしまう可能性がある．特に本研究で対象とする要配慮個人情報は法的にその利用を制限され，可能な限り収集結果から除外するような対策が求められる．

個人情報の保護に関する法律（平成十五年法律第五十七号）[1] (以下，個人情報保護法と略称) では，要配慮個人情報を以下のように定義している．

> 「要配慮個人情報」とは、本人の人種、信条、社会的身分、病歴、犯罪の経歴、犯罪により害を被った事実その他本人に対する不当な差別、偏見その他の不利益が生じないようにその取扱いに特に配慮を要するものとして政令で定める記述等が含まれる個人情報をいう。

学術研究機関等による学術研究目的などいくつかの例外を除き，要配慮個人情報の取得や第三者提供には，原則として本人の同意が必要である．

また，令和 5 年 6 月 2 日には個人情報保護委員会から生成 AI サービスの利用に関する注意喚起等と OpenAI に対する注意喚起 [2] が行われた．その中で，機械学習のための学習用データセットに要配慮個人情報が含まれないように取り組み，もし含まれている場合は削除もしくは適切な加工をすることが要求されている．このような背景から，クローリングにより収集されたコーパスを要配慮個人情報の観点でフィルタリングする必要があり，これを実現するための当該情報の検出・判定技術が不可欠である．

クローリングによるコーパスは非常に大規模であり，特に LLM の事前学習に用いるものでは数兆単語にのぼるケースもある[1)]．このため，判定器は可能な限り高速に動作する必要がある．また，判定器の学習に必要となる要配慮個人情報自体が法律上利用を制限されるため，公開されたデータセットは存在せず，法律に注意しながら独自に構築する必要がある．要配慮個人情報も一種のラベル情報であるため，データセットの構築には既存コーパスに対して各種アノテーション手法を適用することになる．このときに選択する手法によってそれぞれ精度，コスト，時間，労力において異なる特徴を持つ．

本研究では，人手アノテーションと比較して低コストで実行可能な LLM によるアノテーションをコーパスに適用し，要配慮個人情報データセットを構築する．構築したデータセットで各種の機械学習モデルを学習することにより要配慮個人情報の高速な判定器を作成し，それらの性能を比較する．

なお，日本国外でも個人情報保護に関する各種の法律が定められており，コーパスから個人情報を除去することが求められている．Roy ら [3] は PII (個人情報) を検出するために機械学習モデルを構築し，高精度で PII を検出する手法を提案している．

1) https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3

また，LLM の事前学習用コーパスから個人情報を除去する取り組みも進められている．Llama 3 [4] や Gemma 2 [5] では，事前学習用コーパスに対して個人情報やセンシティブな情報を除去するフィルターを適用している．要配慮個人情報の検出にも類似の技術を使用できるが，前節で述べたようにデータの利用方法に制約を受ける点は注意が必要である．

2 要配慮個人情報データセット

この節では要配慮個人情報判定器のためのデータセット構築方法について述べる．ここで構築するデータセットは，判定対象となる文章，および文章に対して付与された 1 個の要配慮個人情報ラベルの組から構成される．

2.1 要配慮個人情報ラベル

個人情報保護法と個人情報の保護に関する法律施行令（平成十五年政令第五百七号）[6] において，1. 人種，2. 信条，3. 社会的身分，4. 病歴，5. 犯罪歴，6. 犯罪被害，7. 障害，8. 診断結果，9. 医療処置，10. 刑事手続，11. 少年保護手続の 11 項目を要配慮個人情報のカテゴリとして定義している．これらの項目に関する詳細は付録に記載する．本研究ではこの 11 項目に加えて，LGBT も要配慮個人情報に含めるか議論されている現状を踏まえて 12. LGBT を追加し，更に 0. 要配慮個人情報以外を加えた 13 項目をラベルとして定義する．

2.2 データセット構築手法

データセットの構築は人名認識と要配慮個人情報アノテーションの 2 段階で行う．まず人名認識を行う理由は，要配慮個人情報は個人情報の一種であるため，要配慮個人情報を含む文章は基本的に人名を含むと推定できることによる．

要配慮個人情報アノテーションには LLM を利用する．ここで，最初から高性能なモデルを使用すると時間やコストが膨大になるため，まず低コストなモデルで予備的なアノテーションを行い，ここで要配慮個人情報を含むと判定された文章に対して，高性能なモデルで最終的なアノテーションを行う．アノテーションには文章が要配慮個人情報を含むかを判定させるプロンプトを使用する．このプロンプトは指示，要配慮個人情報の定義，アノテーション対象の文章で構成され，文章に対応する要配慮個人情報ラベルを出力させる．また，文章が要配慮個人情報を含むと判定された場合は，0 から 1 の範囲で主観的な確信度を出力させる．プロンプトの詳細は付録に記載する．なお，要配慮個人情報ラベルと確信度以外を含む不適切な出力がされた場合は，低パラメータのモデルによるアノテーションでは要配慮個人情報以外として扱い，不適切な出力の少ない高性能なモデルでは人手アノテーションを行う．

表 1 LLM による要配慮個人情報抽出結果

モデル	抽出数	適合率
Swallow	9,220	0.32
Gemma	72,262	0.11
Gemma (確信度 0.8 以上)	15,722	0.16
Swallow+Gemma (確信度 0.8 以上)	2,373	0.55

2.3 データセット構築実験

要配慮個人情報を含む文章を抽出するコーパスは，Common Crawl から収集された日本語の Web コーパスである llm-jp-corpus-v3/ja/ja_cc/level0 [2] (以下，コーパスと略称) の 261 万件の文章を使用した．

人名認識には固有表現認識モデルの bert-base-japanese-v3-ner-wikipedia-dataset [3] (以下，固有表現認識モデルと略称) を使用した．この結果，106 万件の文章に人名が検出された．

日本語に対応したオープンウェイト LLM のうち，低パラメータなモデルでは Llama 3.1 Swallow 8B Instruct v0.2 [7, 8] (以下，Swallow と略称) と Gemma 2 9B IT [9] (以下，Gemma と略称) が優れているとされる [10]．これらのモデルを用いて人名が検出された文章に対するアノテーションを行い，検出された要配慮個人情報を含む文章の件数と，検出結果 100 件の人手確認を基に算出した適合率を表 1 に示す．なお，Swallow+Gemma (確信度 0.8 以上) は両モデルの共通部分を採用している．ここでは，何らかの要配慮個人情報を含む文章を正例，含まない文章を負例として適合率を定義する．Swallow+Gemma (確信度 0.8 以上) の 2,373 件の適合率が最も高く，これを低コストなモデルの結果として採用する．

高性能なモデルは GPT-4o を使用した．GPT-4o を上記の 2,373 件の文書に適用したところ，100 件の人手確認を基に算出した適合率は 0.83 であった．不適切な出力に対して人手でアノテーションした結果を含め，1,087 件が要配慮個人情報を含む文章であ

2) https://gitlab.llm-jp.nii.ac.jp/datasets/llm-jp-corpus-v3/-/tree/main/ja/ja_cc/level0
3) https://huggingface.co/llm-book/bert-base-japanese-v3-ner-wikipedia-dataset

表 2　要配慮個人情報ラベルの内訳

人種	信条	社会的身分	病歴	犯罪歴	犯罪被害	障害	診断結果	医療処置	刑事手続	少年保護手続	LGBT
123	17	6	312	49	20	400	2	7	88	1	62

ると判定された．ラベルの内訳を表 2 に示す．

2.4　データセットの後処理と分析

表 2 より，100 件を超えて収集できたラベルは人種，病歴，障害のみである．ただし，人種は要配慮個人情報と無関係なものを多く含み，このままラベルとして使用するのは適切ではないと考えられた．犯罪歴，犯罪被害，刑事手続は重複して該当する場合があり正確な分類が難しいため，犯罪関係としてまとめることとした結果，これらの合計で 157 件となった．信条，社会的身分，診断結果，医療処置，少年保護手続，LGBT は本手法では十分なデータが確保できなかった．

以上の結果より，以降の節で述べる判定器の学習に用いる正例データとして，本研究では病歴，犯罪関係，障害の合計 869 件を採用した．負例データは図 1 のように，Swallow と Gemma の両方で要配慮個人情報を含まないと判定された文章を使用する．要配慮個人情報を含む文章がコーパス全体に占める割合は非常に低く，判定器が負例優位に学習される可能性が高い．そこで，表 3 に示すように正例データが学習用データセットの 1/10 になるように調整した．

人種，信条，社会的身分は今後データを収集する予定である．なお，診断結果，医療処置は基本的に医療機関が保有する情報であり，Web から収集した情報には含まれないと考えられる．また少年保護手続もデータの性質上，基本的に個人情報が公開されないため，Web 上にはほとんど存在しないと考えられ，存在する場合も犯罪関係に含まれると予想される．

3　要配慮個人情報判定器

3.1　判定器の構築手法

本節で述べる要配慮個人情報判定器は，入力された文章を病歴，犯罪関係，障害，要配慮個人情報以外のいずれかに分類する．具体的には，文章を gensim [11] の doc2vec で埋め込みベクトル化し，これを用いて機械学習モデルで判定する．モデルの学習には前節で述べた要配慮個人情報データセットを用いる．

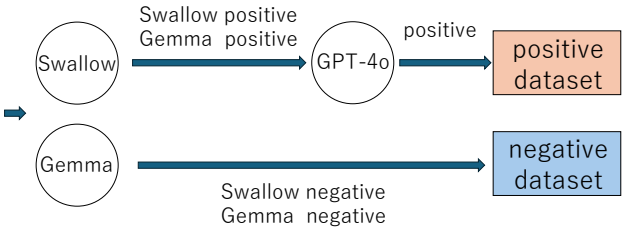

図 1　人名認識後の要配慮個人情報アノテーション

表 3　学習用データセットのラベル内訳

病歴	犯罪関係	障害	要配慮個人情報以外
312	157	400	7,821

3.2　構築実験

doc2vec の学習にはコーパスから固有表現認識モデルで人名を含むと判定されたもののうち 10 万件を使用し，判定器の学習には表 3 の 8,690 件を使用した．機械学習モデルはサポートベクターマシン (以下，SVM と略称)，勾配ブースティング決定木 (以下，GBDT と略称)，ニューラルネットワーク (以下，NN と略称) を使用し，10 エポック学習させた．図 2 に学習データの 5 分割交差検証の結果を示す．クラス重みが None の場合は特別な補正を行わず，balanced の場合はクラス不均衡の補正を自動で行った．ここでも，要配慮個人情報を含む文章を正例，含まない文章を負例として適合率と再現率を定義する．また，F1 スコアは適合率・再現率の各平均値から算出した．

SVM は線形カーネルを使用した場合より RBF カーネルを使用した方が適合率が高くなった．また，NN は doc2vec の表現力を向上させるとクラス重みが None の場合にすべて「要配慮個人情報以外」と判定された．

抽出結果を詳しく分析するために，要配慮個人情報を含む文章を「yes」，要配慮個人情報を含まないが，医療や障害福祉など病歴，犯罪関係，障害のジャンルと少しでも関連性がある情報を含む文章を「maybe」，要配慮個人情報やジャンルと関連性がある情報も全く含まない文章を「no」と定義する．各モデルから得られた要配慮個人情報を含む文章の

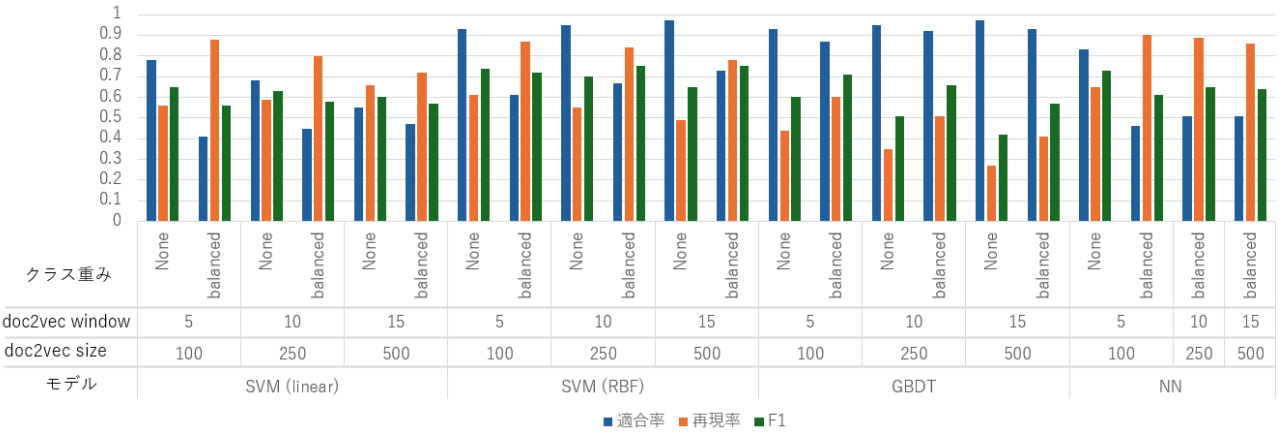

図 2　各判定器の学習データ上の交差検証結果

表 4　判定器による要配慮個人情報抽出結果

モデル	doc2vec size	doc2vec window	yes	maybe	no	抽出数
SVM (RBF & None)	100	5	0.24	0.63	0.13	12,043
	250	10	0.29	0.66	0.05	9,038
	500	15	0.35	0.62	0.03	6,731
SVM (RBF & balanced)	100	5	0.05	0.49	0.46	88,527
	250	10	0.08	0.63	0.29	68,280
	500	15	0.11	0.60	0.29	55,744
GBDT (None)	100	5	0.28	0.57	0.15	8,800
	250	10	0.34	0.55	0.11	4,899
	500	15	0.30	0.64	0.06	3,627
GBDT (balanced)	100	5	0.17	0.71	0.12	22,186
	250	10	0.22	0.63	0.14	12,824
	500	15	0.19	0.72	0.09	9,716
NN (None)	100	5	0.24	0.56	0.20	23,304
NN (balanced)	100	5	0.06	0.49	0.45	118,521
	250	10	0.06	0.54	0.40	121,714
	500	15	0.06	0.53	0.41	120,020

抽出件数と，100 件の人手確認を基に算出した yes, maybe, no の割合を表 4 に示す．

各モデルによって抽出数は大きく異なり，GBDT は抽出数と yes の割合から SVM や NN に比べて再現率は低いことがわかるが，クラス重みを balanced に変更しても適合率があまり低下しないという特徴があった．実際のデータでも全体的にクラス重みを balanced にすると適合率が低下して再現率が向上する傾向があり，doc2vec の表現力を向上させると再現率の変化はモデルによるが，適合率が上昇する傾向があった．また，どのモデルでも図 2 の適合率と比較すると，yes の割合が低下していることが確認できる．

3.3　考察

図 2 について，SVM の RBF カーネルを使用した方が，線形カーネルを使用した場合より性能が優れていたため，データの線形分離が難しいと予想される．また，表 4 と合わせて，クラス重みと doc2vec の表現力を変化させることで，必要に応じて適合率と再現率を調整できることがわかった．

また，yes の割合が図 2 の適合率と比較して低下している原因は，現状の判定器ではジャンルを正しく判定できるが，要配慮個人情報の判定まではできないことにあると考えられる．実際に，表 4 の yes と maybe を足した割合と図 2 の適合率はおおよそ一致しており，ジャンルに関する適合率は図 2 と同様に高い精度を示していると考えられる．このことから，この判定器を 1 段階目で適用してジャンルを分類し，2 段階目で yes と maybe を分類する判定器を適用すれば，より精度の高い判定が実現できると予想される．

4　おわりに

本研究では要配慮個人情報データセットと，要配慮個人情報判定器を構築し性能を確認した．判定器はジャンル分類において高い精度を示したため，要配慮個人情報 (yes) とそれ以外の類似の情報 (maybe) を混合した 1 段階目のフィルタリングに有用であることが示された．また，LLM による判定と比較して低コストで動作するため，大規模なコーパスを処理する際に時間やコストの面で優位性を発揮する．本研究では主に自動的なアノテーション手法について検証したが，今後は人手によるアノテーションの検証も含め，要配慮個人情報に関する判定手法の高性能化に取り組む．

参考文献

[1] デジタル庁. 個人情報の保護に関する法律（平成十五年法律第五十七号）. e-Gov 法令検索. アクセス日: 2024 年 10 月 4 日. https://laws.e-gov.go.jp/law/415AC0000000057/.

[2] 個人情報保護委員会. 生成 AI サービスの利用に関する注意喚起等について. 個人情報保護委員会. アクセス日: 2024 年 10 月 4 日. https://www.ppc.go.jp/news/careful_information/230602_AI_utilize_alert/.

[3] Soumit Roy and Mainak Mitra. Identification and processing of pii data, applying deep learning models with improved accuracy and efficiency. **Journal of Data Acquisition and Processing**, Vol. 33, pp. 1337–1347, 2018.

[4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. **arXiv preprint arXiv:2407.21783**, 2024.

[5] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. **arXiv preprint arXiv:2408.00118**, 2024.

[6] デジタル庁. 個人情報の保護に関する法律施行令（平成十五年政令第五百七号）. e-Gov 法令検索. アクセス日: 2024 年 10 月 25 日. https://laws.e-gov.go.jp/law/415CO0000000507/.

[7] Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[8] Naoaki Okazaki, Kakeru Hattori, Hirai Shota, Hiroki Iida, Masanari Ohi, Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Rio Yokota, and Sakae Mizuki. Building a large japanese web corpus for large language models. In **Proceedings of the First Conference on Language Modeling**, COLM, p. (to appear), University of Pennsylvania, USA, October 2024.

[9] Gemma Team. Gemma. Kaggle, 2024. https://www.kaggle.com/m/3301. DOI: 10.34740/KAGGLE/M/3301.

[10] Swallow LLM. Llama 3.1 swallow. Swallow LLM. アクセス日: 2024 年 11 月 22 日. https://swallow-llm.github.io/llama3.1-swallow.ja.html.

[11] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In **Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks**, pp. 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[12] 個人情報保護委員会. 個人情報の保護に関する法律についてのガイドライン（通則編）. 個人情報保護委員会. アクセス日: 2024 年 10 月 25 日. https://www.ppc.go.jp/personalinfo/legal/guidelines_tsusoku/.

A　要配慮個人情報の定義の補足

要配慮個人情報の各項目は個人情報の保護に関する法律についてのガイドライン（通則編）[12] において詳細に定義されている．以下に要約を示す．

「要配慮個人情報」とは、不当な差別や偏見、その他の不利益が生じないようにその取扱いに特に配慮を要するものとして次の 1 から 11 までの記述等が含まれる個人情報をいう。ただし、1 から 11 の情報を推知させる情報にすぎないもの（例：宗教に関する書籍の購買や貸出しに係る情報等）は、要配慮個人情報には該当しない。

1. 人種　人種、世系、民族的・種族的出身を広く意味する。ただし、国籍や肌の色は含まない。
2. 信条　個人の基本的なものの見方、考え方を意味し、思想と信仰の双方を含む。
3. 社会的身分　個人にその境遇として固着し、一生の間、自力では容易に脱しえない地位を意味し、単なる職業的地位や学歴は含まない。(同和地区出身であることなどが該当する。)
4. 病歴　特定の病気（例：がん、統合失調症など）に罹患した経歴。
5. 犯罪歴　前科、すなわち有罪判決を受けこれが確定した事実。
6. 犯罪被害　犯罪により受けた身体的、精神的及び金銭的被害を意味する。
7. 障害　身体障害、知的障害、精神障害、発達障害、特殊な疾病による障害に関する情報。
8. 診断結果　医師等により行われた疾病の予防、早期発見のための健康診断などの検査結果。
9. 医療処置　医師等による心身の状態の改善のための指導、診療、調剤の情報。
10. 刑事手続　本人を被疑者または被告人として刑事事件に関する手続が行われたという事実。
11. 少年保護手続　非行少年として少年の保護事件に関する手続が行われたという事実。

なお，個人情報は生存している個人に関する情報であり，要配慮個人情報は個人情報の一種であるため，死者に関する情報は要配慮個人情報に該当しない．

B　プロンプトの補足

B.1　Swallow と Gemma のプロンプト

低パラメータのモデルでは 1 回の処理でアノテーションを実行すると精度が低下する傾向が見られた．この問題に対処するため，本研究ではアノテーションを 3 段階で実行した．具体的には，要配慮個人情報の有無の判定，要配慮個人情報に関する人名の有無の判定，生存している個人を特定可能な氏名の有無の判定の 3 段階で順次処理を行った．この 3 段階に分割した理由は，要配慮個人情報は個人情報の一種であるため，人名を含むと推定できること，そして A 章で述べたように，要配慮個人情報は生存している個人に関する情報であるためである．以下に，3 つのプロンプトの指示部分を示す．なお，2 段階目，3 段階目では 1 が出力されたデータを採用する．

1. 入力文書を解析し、要配慮個人情報が含まれるか判断してください
-要配慮個人情報が含まれている場合、その情報が該当するカテゴリのラベルを出力し、判断の確信度も併せて出力してください
-要配慮個人情報が含まれていない場合、0 のみを出力してください

2. 入力文書を解析し、要配慮個人情報に結び付く人名が含まれるか判断してください
-要配慮個人情報に結び付く人名が含まれている場合、1 のみを出力してください
-要配慮個人情報に結び付く人名が含まれていない場合、0 のみを出力してください
-要配慮個人情報が含まれていない場合も 0 のみを出力してください

3. 以下の文書に「生存している個人を特定可能な氏名」が含まれている場合は 1 を、含まれていない場合は 0 を出力してください
注意:
-氏名とは「姓と名が揃ってた形のみ」を指します
-芸名やペンネームでも個人を特定可能なら氏名です
-姓だけ、または名だけしかない場合は氏名ではありません
-架空の人物、故人の名前は氏名ではありません
-英字化、カナ化された仮名は氏名ではありません

B.2　GPT-4o のプロンプト

GPT-4o は 1 回の処理でアノテーションを実行しても精度は高かった．そのため，アノテーションを 1 回の処理で実行した．また，このモデルは多くの指示を与えても正しく処理できる能力を備えている．この特性を活かし，本プロンプトには B.1 節で提示したプロンプトと比較して，要配慮個人情報の対象となる人物が関係者にすぎないかを判定する指示を追加している．以下に，プロンプトの指示部分を示す．

ステップ 1: 入力文書を解析し、要配慮個人情報が含まれるか判断してください
-要配慮個人情報が含まれている場合、ステップ 2 に進んでください
-要配慮個人情報が含まれていない場合、0 のみを出力してください
ステップ 2: 要配慮個人情報の対象である人物のフルネームが含まれるか判断してください
注意: 対象が死者なら 0 を出力してください。また、映画や小説の登場人物でも 0 を出力してください
注意: フルネームでなくイニシャルや名前だけなら 0 を出力してください
-要配慮個人情報の対象となる人物のフルネームが含まれている場合、ステップ 3 に進んでください
-要配慮個人情報の対象となる人物のフルネームが含まれていない場合、0 のみを出力してください
ステップ 3: 要配慮個人情報の対象となる人物が要配慮個人情報の当事者でない関係者 (医者、教授、障害者支援者など) か確認してください
注意: 要配慮個人情報の対象が医者、教授、障害者支援者なら 0 を出力してください
-関係者に過ぎない場合は 0 のみを出力してください
-関係者でなく当事者の場合は、その情報が該当するカテゴリのラベルを出力し、判断の確信度も併せて出力してください