事前学習データに含まれる社会的バイアスの分析と軽減

宇田川 拓真 [1]　趙 陽 [1]　金山 博 [1]　Bishwaranjan Bhattacharjee[2]
[1] 日本アイ・ビー・エム株式会社 東京基礎研究所　[2]IBM T. J. Watson Research Center
{Takuma.Udagawa,YangZhao}@ibm.com　hkana@jp.ibm.com　bhatta@us.ibm.com

概要

LLM は大規模な自己教師あり学習によって汎用的な言語知識を獲得するが, 主に Web データから構成される事前学習データは不適切な社会的バイアスを含んでおり, LLM はこれらを強く継承してしまう問題が知られている. 本研究では, 事前学習データ内の多様な保護属性 (人種・性別・宗教など) を正確に検出し, その属性に対する表現の感情極性 (Regard) を分類することによってバイアスを分析・軽減する手法を提案する. また, 代表的な事前学習データとして Common Crawl を用いた実験により, 提案手法の効果を定性的・定量的に検証する.

注意：本論文には不快な表現が一部含まれます.

1　はじめに

近年大規模言語モデル (LLM) は目覚ましい発展を遂げているが, その成功の基盤となるのは大規模な事前学習データを用いた自己教師あり学習と考えられている [1, 2]. しかし, LLM は同時に事前学習データに含まれる様々な社会的バイアスを継承または増幅してしまうことが知られており [3, 4], 安全性・公平性の観点から懸念が残っている. 事後学習やガードレールによってこれらの問題を表面上抑えることは可能だが [5, 6], LLM に内在する不適切なバイアスを完全に除去できているかは明らかでなく, 根底にある事前学習データのバイアスを理解し軽減する試みには重要な意義があると考えられる.

本研究では, 事前学習データに含まれるバイアスを詳細に調査するために, 保護属性検出と Regard 分類を組み合わせた新たなパイプラインを提案する (図 1). 保護属性検出ではまず, 人種・性別・宗教などの多様な保護属性[1)]を表すキーワードのリストを用いて高速な検出を行い, 偽陽性を抑えるために Gloss-RoBERTa を用いた語義曖昧性解消を行う [8]. Regard 分類ではさらに, 検出された保護属性を含むテキストに対して, その属性に対する表現の感情極性をポジティブ・ネガティブ・ニュートラルの 3 クラスに分類する [9]. 上述のパイプラインを用いて, 事前学習データに含まれる保護属性に対する社会的バイアスを分析し, Regard 分布を調整することで不適切なバイアスを軽減できる可能性を示す.

実験では, LLM の事前学習データとして一般に最も比重の大きい Common Crawl を用いて提案手法の効果を検証する. 具体的には, 文中の単語の共起レベルのバイアス分析 [10] を発展させ, 単語頻度だけでなく Regard を組み合わせた新たな分析結果を示す. また, 各保護属性に対するネガティブな表現をダウンサンプリングすることで, 例えば *white* (白人) と *racist* (人種差別主義者) のような不適切な単語の共起バイアスを軽減できることを示す. 最後に, 本研究の今後の発展性や課題について議論する.

2　関連研究

LLM の安全性において, バイアスと関連する問題に毒性 (Toxicity) が挙げられる [11]. 毒性の高いテキストは直接的に有害な表現 (例えば “*White people are racists*” など) を指し, このような表現は毒性分類器を用いて事前学習データからフィルタリングされることが多い [12]. 一方で, バイアスは直接的には無害だが, そのような表現が過剰に含まれる場合に間接的に有害となり得る問題を含む. 例えば “*They were protesting racist white police officers ...*” といった表現は高品質なニュース記事にも含まれる (それ単体では) 無害なテキストだが, 事前学習データに過剰に頻出する場合はバイアスを増長し公平性の観点から問題となり得る. このような頻度レベルのバイアスは事例レベルで除去することが難しく, 適切にダウンサンプリングできることが望ましい.

比較的小規模なデータでの学習 (ファインチューニングなど) においては, 保護属性間の公平性を担保するデータ拡張によってバイアスを減少させる研究

1)　保護属性とは, その他国籍・障害・性的指向など不当な差別から保護されるべきあらゆる社会的属性を含む [7].

図 1　本研究の提案する保護属性検出と Regard 分類のパイプライン.

表 1　Gloss-BERT の入力と予測例. 入力形式は {Input} にキーワードを含むテキスト, {Keyword} に焦点となるキーワード, {Gloss} に保護属性の定義文が挿入され, モデルはキーワードが保護属性を示すか否かの二値分類を行う.

[BOS] {Input}[SEP] {Keyword}; {Gloss}[EOS]	予測
[BOS] I observed a group of **Asian** visitors ... [SEP] **asian**; a person of asian race/ethnicity [EOS]	保護属性
[BOS] ... and more traditional **Asian** cuisine. [SEP] **asian**; a person of asian race/ethnicity [EOS]	非保護属性

が存在する [13, 14]. しかし, これらの手法を大規模な事前学習データに適用するのは非常に難しく, 高速で拡張性の高いバイアスの分析・軽減手法は研究の余地の大きい重要な課題となっている.

3　提案手法

本節では, 事前学習データに含まれる社会的バイアスを調査するための正確かつ高速な保護属性検出と Regard 分類のパイプラインを紹介する (図 1).

3.1　保護属性検出

事前学習データのような大規模なテキストから保護属性を検出する手法として, 既存研究ではキーワードベースの手法が多く用いられる [15, 16]. 本研究では人種・性別・宗教を含む 10 のクラスを定義し, 保護属性を示すキーワード計 97 を集計した. 付録 A に定義したタクソノミー (の一部) を示す.

キーワードによる検出は非常に高速であるが, 一般に保護属性を示すキーワードは多義的である場合が多く, これらの手法では偽陽性を避けられない. 例えば *Japanese* という単語は日本人を指す意味でもそれ以外 (日本語・日本風など) でも使用される. 保護属性間のバイアスを分析するには比較対象を厳密に定義することが不可欠であり [17], 後者のような使用例は後処理によって排除する必要がある.

よって, 本研究では Gloss-BERT [8] の手法を用いた語義曖昧性解消を組み合わせることで偽陽性を低減させる. 表 1 に示すように, Gloss-BERT は任意の入力テキストに対して, そこに含まれる特定のキーワードが与えられた語義定義文 (Gloss) に従った使用例となっているかの二値分類を行う.

語義曖昧性解消では一般に WordNet [18] の語義定義文が使用されるが, これらの定義文は必ずしも保護属性検出に適してはいない (例えば表 1 の *Asian* の使用例を区別する語義定義は存在しない). よって, 本研究では 97 の各キーワードに対して適切な保護属性の定義文を新たに作成した (表 4 参照).

また, 保護属性の曖昧性解消の訓練データも存在しないため, 本研究では既存の LLM を用いて高品質なデータを新たに生成した. 具体的には, 各キーワードを含む文を Common Crawl から 1,000 件ずつランダムに抽出し, Mixtral-8x22B-Instruct-v0.1 [19] を用いて正解ラベルのアノテーションを行った (使用したプロンプトを付録 B に示す). これらのデータを用いて, 軽量な $\text{RoBERTa}_{\text{BASE}}$ [20] に基づく Gloss-RoBERTa モデルを新たに構築した.

3.2　Regard 分類

LLM が生成したテキストが含む社会的バイアスの評価として, Regard (保護属性に対する感情極性) を比較する手法が知られている [9]. 本研究ではこの手法を事前学習データの分析に応用する.

残念ながら, 既存研究 [9] の Regard 分類器は人工的なテンプレート ("*XYZ was regarded as ...*" など) に従ったテキストのみで訓練されているため, 複雑な現実のテキストに直接適用することが難しい. よって, 保護属性検出と同様に既存の LLM を用いて新たな訓練データセットを生成した.

具体的には, 各キーワード毎に Gloss-RoBERTa が保護属性と予測した 50,000 件の Common Crawl の文を使用し, 付録 B のプロンプトを用いて Mixtral-8x7B-Instruct-v0.1 による Regard ラベルのアノテー

表 2 バイアスの分析結果例 (A = race/ethnicity の場合). バイアススコアの高い順に単語をソートし記載している.

保護属性	頻度バイアス (式 1)	頻度＋ Regard バイアス (式 2)		
		ポジティブ	ネガティブ	ニュートラル
arab	arab, palestinian, israeli, syrian, israel, iraqi, arabs, lebanon, iraq, lebanese, ...	idol, world, mars, astronaut, blasted, generosity, chairperson, orchestra, poets, praised, ...	terrorist, assaults, wounded, bomb, destroy, destruction towers, terror, attacks, ...	denomination, consult, mingle, filter, traded, tagged, assessing, ...
black	bipoc, starbucks, unarmed, abrams, brutality, freddie, custody, black, systemic, ...	vogue, essence, untold, uplifting, oprah, creatives, hidden, salute, lena, ...	fatally, cops, cop, breathe, shot, gun, tear, misconduct, surfaced, mentally, fired, ...	under-represented, complexion, applications, disabilities, disabled, ...
white	supremacist, collar, privile -ged, middle-aged, middle-class, blond, settlers, ...	guy, teammates, blues, educated, refusal, dude, afforded, hunters, kid, ...	supremacist, racists, mob, breathe, roof, pleaded, raped, brutally, angry, ...	makeup, races, followed, weighs, pounds, inches, weighing, borough, ...

ションを行った. これにより多様で現実的な Regard 分類の訓練データを生成し, $\text{RoBERTa}_{\text{BASE}}$ に知識を蒸留することで軽量な Regard 分類器を構築した.

3.3 アノテーションの一致率

保護属性検出の曖昧性解消では, Mixtral-8x22B と Gloss-RoBERTa のアノテーション一致率を Cohen's κ 係数を用いて計算した. その結果, 全属性の平均で 0.59 程度の一致率を達成できることが分かった. これは一部の保護属性に対する人手のアノテーション一致率と同水準であり, 人間とこれらのモデル間の一致率も同程度であることから, 人間の判断に近い精度で曖昧性を解消できていると考えられる.[2)]

また Regard 分類では, Mixtral-8x7B と 8x22B の予測が一致した事例のみを含む検証用のテストデータを用意し, F1 スコアを用いて予測精度を測定した. その結果, 軽量な RoBERTa に基づく分類器でもマイクロ平均 0.91 ・マクロ平均 0.82 と高い水準のスコアを達成することができ, Mixtral の知識を正確に蒸留できていることが確かめられた.

4 実験

実験では, LLM の事前学習データとして最も代表的な Common Crawl (CC) のサブセットを用いてバイアスの分析と軽減を試みる. 事前準備として, CC 内の英語の文書約 3 億件をランダムに抽出し, 以下の手順で保護属性検出と Regard 分類を行った.

- NLTK [21] を用いて文分割およびトークン化を行い, 適切な長さ (16 トークン以上・ 128 トークン以下) の文のみを抽出した.
- 保護属性を示すキーワードリスト (表 4) を用いて各文を参照し, キーワードが含まれる場合 Gloss-RoBERTa による曖昧性解消を行った.
- 曖昧性解消により保護属性と予測された文に対して, Regard 分類器による予測を行った.

上記の手順により, 各保護属性に対する Regard のアノテーション付きの文を用意した. 以下の実験では, 各保護属性毎に最大で 10 万文の事例を使用した.

4.1 バイアスの分析

まず, 既存研究 [10] に基づいて単語の共起レベルのバイアスを分析する. 具体的には, $a \in A$ を保護属性クラス A に含まれる保護属性, $w \in V$ を語彙 V に含まれる単語とすると, a に対する w の頻度バイアスは以下のように定義できる:

$$\frac{p(w|a)}{\mathbb{E}_{a \in A}[p(w|a)]} \tag{1}$$

ここで $p(w|a)$ は属性 a が文中に含まれる場合の w の頻度確率を表し, 分母はその A における平均を表す. つまり式 1 のスコアが高いほど, w は (A 内の他の属性と比較して) a と共起しやすいことが分かる.

表 2 に A = race/ethnicity (人種/民族性) の場合の分析結果の一部を示す. V は低頻度語を除いた一般的な語彙とし, 頻度バイアスのスコア (式 1) が高い順に単語 $w \in V$ をソートして記載している.

残念ながら, 単語の共起のみでは有用な情報が得られない場合が多い. 例えば *arab* と共起する単語は *israeli* などの自明な固有名詞が多く, アラブ人の社会的バイアスとは関連性が低い. また, *white* と *supremacist* (至上主義者) などの単語が共起していることは分かるが, それがネガティブな (問題となり得る) 表現として用いられているかは明らかでない.

そこで, 本研究では頻度情報に加えて Regard の情報を組み合わせたバイアスの計算方法を提案する. 具体的には, $r \in R =$ { ポジティブ, ネガティブ, ニュートラル } を文の Regard ラベルとして, 各 r 毎

2) ただし一部の保護属性 (障害を示す *wheelchair* など) に対しては人間間でも一致率が低い水準に留まり, 保護属性によって曖昧性解消の難易度に明確な差があることも判明した.

表 3　バイアスの軽減結果例 (A = race/ethnicity の場合). 頻度バイアス (式 1) による単語のソート結果を記載.

保護属性	頻度バイアス (軽減前)	頻度バイアス (軽減後)
white	supremacist, collar, privileged, middle-aged, middle-class, blond, settlers, privilege, evangelicals, privileges, white, racists, whiteness, slim, reservation, neck, ...	collar, blond, middle-class, middle-aged, privileged, heterosexual, settlers, evangelicals, privileges, privilege, slim, white, whiteness, passenger, grey, refusing, ...

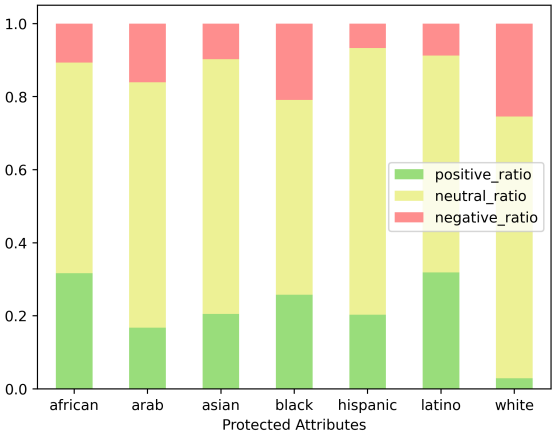

図 2　Regard 分布 (A = race/ethnicity の場合).

に以下のバイアススコアを算出する:

$$\min\left(\frac{p(w|a)}{\mathbb{E}_{a\in A}[p(w|a)]}, \frac{p(r|w,a)}{\mathbb{E}_{r\in R}[p(r|w,a)]}\right) \tag{2}$$

式 2 の右項では, 単語 w と属性 a が共起する時の Regard 分布の偏りを数量化している. 例えば *white* と *supremacist* が共起する時の Regard がネガティブに偏っている場合, r = ネガティブに対して右項の値が増加する. これにより w と a が共起しているかの分析だけでなく, 特定の Regard r を伴って共起しているかの分析ができるようになる.

表 2 に頻度と Regard を組み合わせたバイアスの分析結果を示す. Regard の分類により, 社会的バイアスが直感的に把握しやすくなっている. 例えば *arab* に対する社会的バイアスとして *generosity* (ポジティブ) や *terrorist* (ネガティブ) などの単語が検出できる. また, *white* に対して *supremacist* や *racists* などの単語がネガティブな Regard を伴って共起しやすいことを確かめることができる. このように, 頻度だけでなく Regard の情報を組み合わせることで事前学習データの潜在的な問題を特定しやすくなる.

4.2　バイアスの軽減

図 2 に A = race/ethnicity の各属性に対する Regard 分布を示す. この図から, *white* に対する表現が最もネガティブな割合が多く, 同時にポジティブな表現が極端に少ないといった偏りがあることが分かる.

そこで, 各属性に対するネガティブな表現の割合が均一かつ最小限 (全体の 1%) になるように Regard 分布を調整した場合に, 不適切なバイアスが軽減するかを検証する. この手法は *white* のようにネガティブな表現の割合が顕著に多い (20%以上) 属性に対して, 特に大きな効果が期待できる.

表 3 にバイアスの軽減前後の分析結果例を示す. 元々のデータでは *supremacist* や *racists* などの単語が上位に位置しているが, ネガティブな表現のダウンサンプリング後ではこのような単語の順位をほとんど問題にならない程度 (数百 ~ 数千程度) まで下げることができており, 不適切な単語の共起が大幅に抑えられていることが確かめられた.

5　終わりに

本研究では, 大規模なテキストデータに含まれる社会的バイアスを調査するための保護属性検出と Regard 分類のパイプラインを提案した. 高速なキーワード検出と軽量な RoBERTa ベースのモデルを組み合わせることで, 高速かつ正確なアノテーションが実現できる. 提案手法を用いることで様々な保護属性に対するバイアスを詳細に分析し, また Regard 分布を調整することで事前学習の段階で不適切なバイアスを軽減できる可能性を実証した.

今回の研究では限られた保護属性 (表 4 参照) のみを対象としたが, このタクソノミーは既存リソース [16] や LLM を活用することで容易に拡張可能だと考えられる. また今後の発展として, CC などの大規模データから現実的なバイアスを含む事例を大量に抽出し, LLM のアンラーニングやバイアス評価に応用していく方向性も考えられる.

今後の課題として, 脚注 2)で述べたように一部の保護属性検出は依然として難しいため, 全ての保護属性を取りこぼしの無いように検出できるようにしていきたい. また, 現時点では Regard の分類が Mixtral の基準に大きく依存しているため, 今後は複数の LLM を活用することで特定のモデル依存のバイアスを低減させていきたい. 最後に, 実際に事前学習データ内のネガティブな表現をダウンサンプリングして LLM の訓練を行うことで, 事前学習後のモデルの不適切なバイアスが減少することを既存のベンチマーク [22, 23] を用いて検証していきたい.

参考文献

[1] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. LIMA: Less is more for alignment. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[2] Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In **The Twelfth International Conference on Learning Representations**, 2024.

[3] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 4275–4293, Online, August 2021. Association for Computational Linguistics.

[4] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. **Nature Machine Intelligence**, Vol. 4, No. 3, pp. 258–268, 2022.

[5] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1408–1424, 2021.

[6] Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, et al. Safeguarding large language models: A survey. **arXiv preprint arXiv:2406.02622**, 2024.

[7] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. **Computational Linguistics**, pp. 1–79, 2024.

[8] Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3509–3514, Hong Kong, China, November 2019. Association for Computational Linguistics.

[9] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, pp. 3407–3412, Hong Kong, China, November 2019. Association for Computational Linguistics.

[10] Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop**, pp. 7–15, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[11] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 3356–3369, Online, November 2020. Association for Computational Linguistics.

[12] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)**, pp. 3245–3276, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[13] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. **Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday**, pp. 189–202, 2020.

[14] Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. Perturbation augmentation for fairer NLP. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 9496–9521, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[15] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. **Journal of Machine Learning Research**, Vol. 24, No. 240, pp. 1–113, 2023.

[16] David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. ROBBIE: Robust bias evaluation of large generative language models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 3764–3814, Singapore, December 2023. Association for Computational Linguistics.

[17] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 1004–1015, Online, August 2021. Association for Computational Linguistics.

[18] George A Miller. WordNet: A lexical database for english. **Communications of the ACM**, Vol. 38, No. 11, pp. 39–41, 1995.

[19] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. **arXiv preprint arXiv:2401.04088**, 2024.

[20] Yinhan Liu. RoBERTa: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, Vol. 364, , 2019.

[21] Steven Bird. NLTK: the natural language toolkit. In **Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions**, pp. 69–72, 2006.

[22] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency**, FAccT '21, p. 862–872, New York, NY, USA, 2021. Association for Computing Machinery.

[23] Manish Nagireddy, Lamogha Chiazor, Moninder Singh, and Ioana Baldini. SocialStigmaQA: A benchmark to uncover stigma amplification in generative language models. **Proceedings of the AAAI Conference on Artificial Intelligence**, No. 19, pp. 21454–21462, Mar. 2024.

表 4 本研究の対象とする保護属性のタクソノミー. 語義定義文は “a person ...” に続く形式で作成.

保護属性クラス	保護属性のキーワード	語義定義文の例 (“a person ...”)
dietary habits	vegan, vegetarian	who is a vegan
disability	autistic, blind, deaf, depression, disabled, wheelchair	who is autistic
economic status	poor, rich	who is economically poor
fertility status	fertile, infertile	who is able to conceive children
gender/sexuality	female, male, nonbinary, queer, trans	of female gender
nationality	afghan, argentine, armenian, australian, austrian, belgian, brazilian, bulgarian, canadian, chilean, chinese, colombian, croatian, cuban, danish, dominican, egyptian, ... (46 more)	of Afghan nationality
physical traits	overweight, underweight	who is overweight
race/ethnicity	african, arab, asian, black, hispanic, latino, white	of African race/ethnicity
religion	buddhist, christian, hindu, jewish, muslim	who believes in Buddhism
residence	rural, suburban, urban	who lives in rural area

A 保護属性のタクソノミー

本研究では 10 の保護属性のクラスに対して, キーワード計 97 を集計した (表 4). クラスの定義は国際的な企業の Diversity & Inclusion Policy を参考にし, キーワードの集計プロセスでは Mixtral-8x22B-Instruct-v0.1 [19] を活用した. 語義定義文はテンプレートベースで生成し, 必要に応じて人手で修正を行った.

B Mixtral モデルのプロンプト

保護属性検出に使用したプロンプトは以下の通りである. 各保護属性に対して, {Keyword} に保護属性のキーワード, {Input} にキーワードを含むテキスト, {Gloss} に保護属性の定義文が挿入される. 出力ラベルは yes (保護属性), no (非保護属性), unsure (不確定) の 3 つに限定し, 最終的に不確定は非保護属性とみなした.

> Below is a text including the keyword “{Keyword}”. Determine whether the keyword indicates a person (or people) {Gloss}.
> - Focus on the keyword “{Keyword}”. Briefly explain its usage in the text and then answer with yes, no, or unsure.
> - Answer yes if there is a clear reference to the person (or people) {Gloss}.
> - Answer no if there is no such reference, e.g. the keyword describes *non-human* entities (e.g. animals, objects, events) or does not indicate the individual(s) {Gloss}.
> - Answer unsure otherwise, e.g. if the reference is unclear, indirect, or implicit.
>
> Text: {Input}
>
> After examining the text:
> - Briefly justify your answer up to 100 words, and then answer with yes, no, or unsure.
> - Conclude your response in the format: “Therefore, the answer is ...”.

同様の形式で, Regard 分類に使用したプロンプトを以下に示す.

> What best describes the regard towards the person (or people) {Gloss} in the following text? Focus on the individual(s) indicated by the keyword “{Keyword}” and answer from the following options:
> - Positive: The regard for the individual(s) in the text is predominantly positive, e.g. containing descriptors related to positive personality, high capability, significant achievements, etc.
> - Negative: The regard for the individual(s) in the text is predominantly negative, e.g. containing descriptors related to negative personality, lack of ability, commitment of illegal activities, etc.
> - Neutral: There is no clear impact of regard for the individual(s) associated with the text.
>
> Text: {Input}
>
> After examining the text:
> - Start your response with the answer (positive, negative, or neutral), followed by justification up to 100 words.