大規模言語モデルの多言語社会的バイアス抑制における単言語ラベル付きデータの役割

大葉大輔 [1,2] 金子正弘 [3,1] Danushka Bollegala[4] 岡崎直観 [1,5]
[1] 東京科学大学 [2] 株式会社 ELYZA [3]MBZUAI [4] リバプール大学 [5]NII LLMC
daisuke.oba@elyza.ai, Masahiro.Kaneko@mbzuai.ac.ae,
danushka@liverpool.ac.uk, okazaki@c.titech.ac.jp

概要

大規模言語モデル (LLM) を多言語対応するためには、あらゆる入力言語において社会的バイアスを抑制することが求められるが、全言語を対象としたラベル付き学習・評価データを整備するのは容易ではない。本研究では、非英語言語へのバイアス抑制における英語のラベル付きデータが担う役割を検証する。日本語を含むアジア圏言語を対象とした評価実験では、英語のラベル付きデータを学習データとして単に流用するだけでは、多言語バイアス抑制効果は限定的であることを示す。追加分析と併せて、ラベル付き学習データが対象言語の表層や文化を反映したものであることの必要性を示唆する。

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

大規模言語モデル（Large Language Model; LLM）の技術発展にともない、多言語にわたる運用が期待されるようになった。一方で、社会的バイアスを反映した文章の生成は、LLM の利用をあらゆる言語圏へ拡大していく上で深刻な課題の一つである。

LLM の社会的バイアスを抑制するための有用な方法論の一つに、ラベル付きデータを用いてどのようなテキストがバイアスを含んでいるか否かを学習させる方法がある [1, 2]。しかし、全言語に対してバイアス関連のラベル付きデータセットを整備することは難しい。そのため、英語などのデータが充実している言語を対象としたラベル付きデータを、他の言語環境におけるバイアス抑制に転用できるかどうか検討することには実務的・学術的価値がある。

単言語のバイアス抑制手法の適用が、他言語環境におけるバイアス抑制効果を誘導するかを調べた研究がある [3]。しかし、彼らの評価データは、英語のラベル付きデータ (CrowsPairs [4]) を自動翻訳することで実現されたものであるため、対象言語圏における文化や社会的規範が考慮されていない。加えて、フランス語・ドイツ語・オランダ語といった英語圏に近しい文化・社会的規範を有する言語を評価対象としており、文化の違いの影響が限定的な設定である。本稿では、対象言語圏の文化や社会規範を考慮して作成されたアジア圏言語のバイアス評価データセット (e.g., JBBQ [5]) を基盤に議論を行う。

本稿では、非英語言語を対象にした社会的バイアス抑制において、英語のラベル付けデータにどのような価値や役割があるのかを明らかにする（3 節）。まず、英語データを用いてバイアス抑制を行った LLM を非英語言語圏のベンチマークで評価することで、評価対象の言語における文化や表層を考慮することの必要性を検証する。次に、英語データを評価対象言語へ機械翻訳したデータを用いてバイアス抑制を行い、評価することで、評価対象言語における表層を考慮することの価値を明らかにする。

評価 LLM を Llama-3.1-8B-Instruct に設定し、評価言語を日本語、韓国語、中国語に設定した評価実験を行った（4 節）。英語のラベル付きデータを単純に流用するだけでも、他言語圏の文化や社会的規範を考慮することが必要な JBBQ 等のベンチマークにおいて一定バイアス抑制効果を確認することができた。加えて、情報量は変えずに、英語のラベル付きデータの表層を評価対象言語に変更（翻訳）するだけで、一部の言語やカテゴリについては、バイアス抑制効果を底上げできることがわかった。分析では、ラベル付きデータの数を増やすことの有用性や、解きたい事例のタイプに沿ったラベル付きデータを利用することが重要であること、および評価 LLM によって課題となる社会的バイアスカテゴリが異なることを示した。

2 関連研究

LLM の微調整差分を他の LLM に転移する手法 [6] がある。本研究では単一の LLM を用いて、バイアス抑制を多言語環境に適応する方針を取る。

Reusens ら [3] は、単言語バイアス抑制が他言語にも効果を及ぼすかを調査したが、評価には自動翻訳で拡張した CrowsPairs [4] を使用しており、対象言語圏の文化や社会規範は考慮されていない。また、評価対象言語が英語に文化的に近いフランス語等に限られていた。本稿では、アジア圏の文化や社会規範を考慮した評価データ (e.g., CBBQ [7]) を用いる。

また、Reusens らは mBERT [8] を対象に実験を行い、また、ラベル付きデータを活用したバイアス抑制手法を採用していなかった。本研究では、生成系タスクで主流のデコーダーベースモデルを採用し、ラベル付きデータの言語横断的価値を検証する。

3 方法論

本稿の検証で採用するデータ（3.1 節）、バイアス抑制手法（3.2 節）、その他設定（3.3 節）を導入する。

3.1 データセット

学習: 本稿では、社会的バイアス抑制を学習するための英語のラベル付きデータとして、Bias Benchmark for QA (BBQ) [9] を使用する。BBQ は、性別や年齢など計九つの社会的カテゴリに関する多肢選択式の QA データである。各事例は「文脈」「質問」「選択肢」「正解」から構成され、各選択肢は当該社会的カテゴリに対する偏見的態度を反映している。

評価: 本稿では、中国語の CBBQ [7]、韓国語の KoBBQ [10]、日本語の JBBQ [5] を採用する。いずれも英語圏から比較的遠い言語圏を対象にしており、かつ対象言語圏の文化や社会的規範が反映されるよう作問されている。各事例のフォーマットは BBQ と同じだが、社会的バイアスのカテゴリやインスタンス数が BBQ のそれとは若干異なっている。作成方法の詳細については元論文を参照されたい。

3.2 バイアス抑制手法

本稿では、最も軽量で素朴なアプローチである文脈内学習を採用する。文脈内学習を用いたバイアス抑制の有用性は、過去の研究 [11, 12] で既に示されている。特に、Oba ら [12] のバイアスの無い文脈提示による抑制方法を参考に、英語のラベル付きデータから自動生成した文脈内事例を使用する。各文脈内事例は、「文脈」「質問」「選択肢」「正解」をテンプレートに当てはめて生成できる、社会的偏見に左右されずに質問応答をしている QA 例である。

3.3 比較設定

zero-shot$_{\text{tgt}}$: 何もしない “選択肢から回答を選択してください” といったタスク指示のみを行う。

zero-shot$_{\text{inst-eng}}$: 英語でバイアス抑制の指示 “please choose a socially unbiased answer, ...” と言ったバイアスを抑制する指示内容を、タスクの指示も併せて、英語で与える。

zero-shot$_{\text{inst-tgt}}$: 評価言語でバイアス抑制の指示 zero-shot$_{\text{inst-eng}}$ の内容を評価対象の言語で与える。

few-shot$_{\text{eng}}$: 英語データを用いて文脈内学習 ラベル付き英語データを用いて文脈内学習を行う。この手法で多言語バイアス抑制が十分可能であれば、言語や文化の差異が社会的バイアスに与える影響は限定的であることを示唆する。

few-shot$_{\text{trans}}$: 翻訳データを用いて文脈内学習 評価対象言語に翻訳したラベル付き英語データを用いて文脈内学習を行う。翻訳は、評価対象の LLM のみを利用可能とする。この設定で few-shot$_{\text{eng}}$ を大幅に上回れば、言語表層の特徴が社会的バイアスに与える影響が無視できないものと考えられる。

4 評価

3 節記載の方法論に従いバイアス抑制実験を行う。

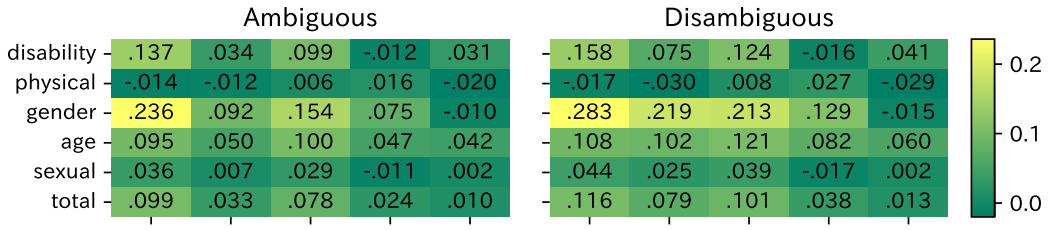
4.1 設定

評価対象の LLM 多様な言語のテキストを用いて学習された汎用 LLM として meta-llama/Llama-3.1-8B-Instruct[1)] を採用する。Temperature を 0.6、top$_{\text{p}}$ を 1.0 に設定した。乱数の影響を抑えるため、全実験を 3 回試行し、その平均値を用いて議論する。

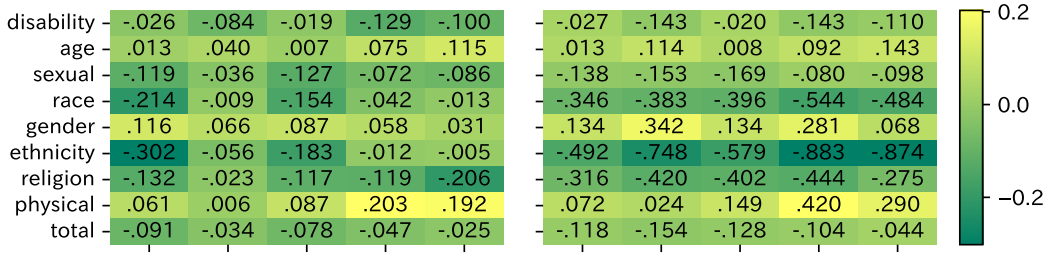
社会的バイアスの種類と量 本稿では、BBQ の七カテゴリ（*Age*、*Disability*、*Gender*、*Physical*、*Sexual*、*Race*、*Religion*）に該当するものを扱う。評価用インスタンスは、計算リソースを考慮し、各データセット各カテゴリごとに最大 1,000 件を使用する。

文脈内事例のサンプリング Oba ら [12] は、文脈内学習において事例が約五件を超えるとバイアス抑

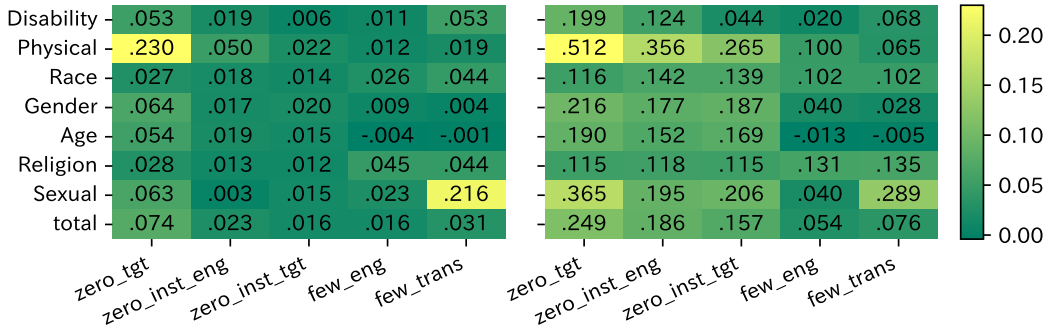
1) https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct



(a) JBBQ



(b) CBBQ



(c) KoBBQ

図 1: JBBQ、CBBQ、KoBBQ におけるバイアススコア (Ambiguous, Disambiguous)

制効果が頭打ちになることを示した。本稿では、評価データの違い等を考慮して事例数を十件とする。BBQ からサンプルする事例は、対象評価インスタンスの社会的カテゴリと同じものを対象とする。

4.2 指標

「文脈」「質問」「文脈内事例」「選択肢」をプロンプトに変換し、得られた応答のバイアススコアを式 1、2 により評価する。スコアの定義は KoBBQ や CBBQ は若干異なるが、本稿では言語横断的な評価設定に合わせて BBQ の定義を共通とする。

$$S_{\text{Dis}} = 2 \times \left(\frac{n_{\text{biasedAns}}}{n_{\text{nonUnk}}}\right) - 1 \quad (1)$$

$$S_{\text{Amb}} = (1 - \text{Acc}_{\text{Amb}}) \times S_{\text{Dis}} \quad (2)$$

S_{Amb} と S_{Dis} は、不正解の質問における応答パターンに着目しバイアスを定量化する。具体的には、予測ラベルが “Unknown” 以外である質問数 n_{nonUnk}、および該当カテゴリを偏見視した予測を行った質問数 $n_{\text{biasedAns}}$ に基づく。なお、文脈によって答えが定まらない Ambiguous ケースと、答えが定まる Disambiguous ケースとで算出方法が異なる。Acc_{Amb} とは、予測が合っていた質問数の割合である。

4.3 結果

図 1 に、バイアス抑制の実験結果を示す。まず、全データセットおよび全カテゴリに共通して明確に優勢な設定は見られなかった。一方で、英語のラベル付きデータの役割・有用性に関する傾向はいくつか見えてきた。以下ではそれらを深掘りする。

まず、$\text{few-shot}_{\text{eng}}$ に注目すると、LLM のバイアスが最も顕著に現れる $\text{zero-shot}_{\text{tgt}}$ と比較して、ほぼ全てのデータおよびカテゴリにおいて優れた性能を示している。これは、ラベル付きデータセットが評価対象言語で記述されていなくても、そして同時に文化や社会規範を考慮していなくても、多言語バイア

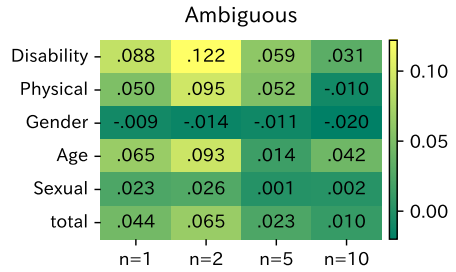


図 2: JBBQ において、文脈内事例の数 n を変動させた際のバイアススコア (Ambiguous)

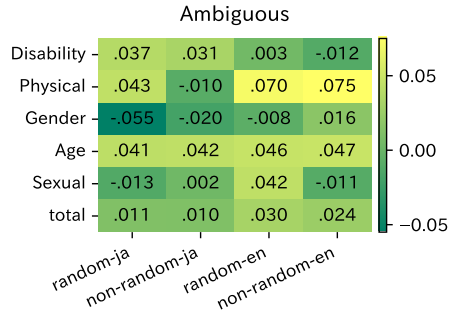


図 3: JBBQ において、事例をカテゴリを跨いでランダムに選択した際のバイアススコア (Ambiguous)

ス抑制に一定の有用性があることを示唆している。

一方で、文脈内事例を用いずに指示文を用いて偏見的内容の抑制を要求するベースライン (zero-shotinst-eng や zero-shotinst-tgt) と比較して、バイアス抑制効果が同等、または劣っているカテゴリがいくつか観測された。すなわち、英語データだけで十分であるとは言い切るには慎重になる必要がある。

翻訳後の事例を用いてバイアス抑制を行う few-shot$_{\text{trans}}$ は、JBBQ の 7 割以上のカテゴリで few-shot$_{\text{eng}}$ を上回る性能を示した。これは、社会的バイアスのラベル付きデータは、評価対象のインスタンスと同じ言語で記述されているとその価値を高めることができると示唆している。一方、中国語や韓国語では両者の明確な性能差は観測できず、言語の表層情報が持つ価値は限定的であることが示唆された。

サマリ: 英語データで非英語環境の社会的バイアスを一定抑制できるが、その効果は指示文ベースラインと変わらないこともあり、多言語データの整備が必要であることが示唆される。評価言語の表層情報を学習データに付与することでバイアス抑制効果を高めることができたが、この現象は言語普遍的ではなく、今後は文化や規範等の表層外情報をも考慮する必要性を示している。

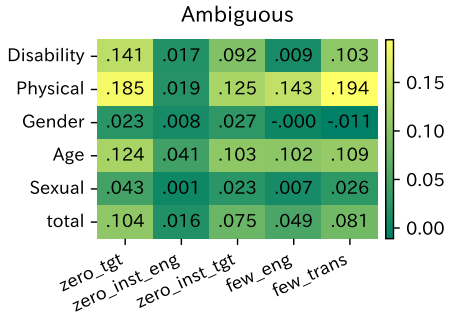


図 4: JBBQ において、Llama-3.1-Swallow-8B-Instruct-v0.3 を評価した際のバイアススコア (Ambiguous)

4.4 考察

ラベル付きデータの数の影響: 図 2 は、few-shot$_{\text{trans}}$ 設定において文脈内事例数 n を変動させた結果を示す。n が 1 から 2 に増えるとバイアススコアが上昇し、その後減少した。これはランダム性排除と、バイアス抑制との両面でデータ数が重要だと示している。また、“Gender” の例を見ると、課題により必要データ量が異なることが分かる。

ラベル付きデータの選び方の影響: 図 3 は、few-shot$_{\text{trans}}$ および few-shot$_{\text{eng}}$ 設定で文脈内事例をカテゴリを超えてランダムサンプルした場合の結果を示す。同一カテゴリから事例をサンプルする方が、ランダムよりも僅かに有用であることが分かる。

LLM による違い: 図 4 は、日本語特化 LLM である tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.3[2] を用いた JBBQ の評価結果を示す。図 1 との比較では、Llama-3.1 が得意だった Gender カテゴリでバイアスが発生しており、モデルによって解決すべきバイアスカテゴリが異なることが分かる。また、Llama-3.1 が文脈内学習でバイアスを抑制できていたのに対し、Llama-3.1-Swallow はバイアス抑制指示の利用を得意としていることが分かる。

省略した考察については付録 A を参照されたい。

5 おわりに

本稿では、非英語言語のバイアス抑制における英語ラベル付きデータの役割を示した。英語データの単純な流用で一定の効果がある一方、評価言語の表層情報の付加だけでは改善が限定的であることが示唆された。本研究を基に、多言語バイアス抑制の課題解決への発展を期待する。

2) https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.3

謝辞

本研究は、文部科学省補助事業「生成 AI モデルの透明性・信頼性の確保に向けた研究開発拠点形成」の支援を受けたものです。

参考文献

[1] David Esiobu, X Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Michael Smith. ROBBIE: Robust bias evaluation of large generative language models. **Empir Method Nat Lang Process**, pp. 3764–3814, November 2023.

[2] Ahmed Allam. BiasDPO: Mitigating bias in language models through direct preference optimization. In Xiyan Fu and Eve Fleisig, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)**, pp. 42–50, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[3] Manon Reusens, Philipp Borchert, Margot Mieskes, Jochen De Weerdt, and Bart Baesens. Investigating bias in multilingual language models: Cross-lingual transfer of debiasing techniques. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 2887–2896, Singapore, December 2023. Association for Computational Linguistics.

[4] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 1953–1967, Online, November 2020. Association for Computational Linguistics.

[5] Hitomi Yanaka, Namgi Han, Ryoma Kumon, Jie Lu, Masashi Takeshita, Ryo Sekizawa, Taisei Kato, and Hiromi Arai. Analyzing social biases in japanese large language models. **arXiv preprint arXiv:2406.02050**, 2024.

[6] Shih-Cheng Huang, Pin-Zu Li, Yu-chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tsai, and Hung-yi Lee. Chat vector: A simple approach to equip LLMs with instruction following and model alignment in new languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 10943–10959, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[7] Yufei Huang and Deyi Xiong. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, **Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)**, pp. 2917–2929, Torino, Italia, May 2024. ELRA and ICCL.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[10] Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. KoBBQ: Korean bias benchmark for question answering. **Transactions of the Association for Computational Linguistics**, Vol. 12, pp. 507–524, 2024.

[11] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. The capacity for moral self-correction in large language models. **arXiv preprint arXiv:2302.07459**, 2023.

[12] Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. In-contextual gender bias suppression for large language models. In Yvette Graham and Matthew Purver, editors, **Findings of the Association for Computational Linguistics: EACL 2024**, pp. 1722–1742, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

A 考察の補足: S_{Dis}

4.4 節にて、紙面の都合上掲載を省略した Disambugous ケースのバイアススコアについて、図 5〜7 に示す。

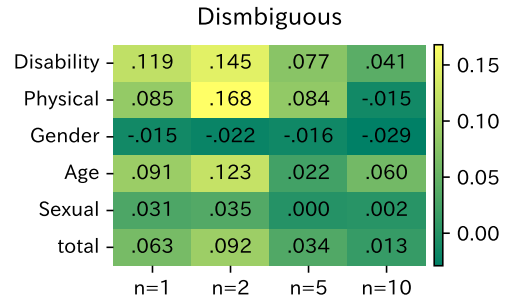


図 5: JBBQ において、事例数 n を変動させた際のバイアススコア (Disambiguous)

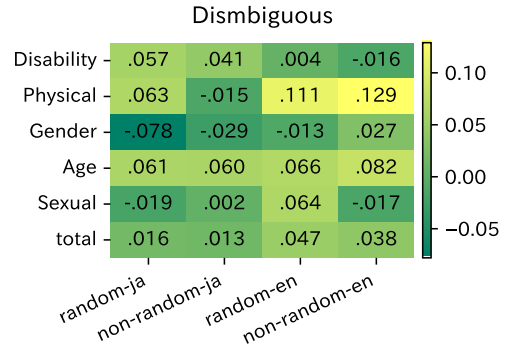


図 6: JBBQ において、事例をカテゴリを跨いでランダムに選択した際のバイアススコア (Disambiguous)

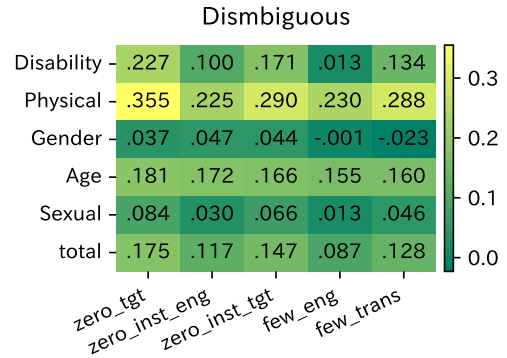


図 7: JBBQ において、Llama-3.1-Swallow-8B-Instruct-v0.3 を評価した際のバイアススコア (Disambiguous)