出力埋め込み操作による統計的および因果的バイアスの抑制

石戸谷由梨 [1]　小林一郎 [1]
[1] お茶の水女子大学大学院
{g2120505,koba}@is.ocha.ac.jp

概要

大規模言語モデル (LLM) におけるバイアスは重要な課題であり，様々なバイアス軽減手法が考案されてきた．その中でも，本研究は統計的公平性を損なう統計的バイアスと，因果的公平性を損なう因果的バイアスという異なる性質をもつ二種類のバイアスに着目する．本研究では，これらのバイアスに対して異なる設計意図をもつ介入を出力埋め込みに限定して適用する枠組みを検討する．実験の結果，因果的バイアスを直接的に軽減することは確認されなかった一方で，因果的バイアスを対象とした介入が，統計的バイアスの軽減に寄与することが観測された．この結果は，統計的バイアスと因果的バイアスを独立に扱うことの難しさを示唆しており，LLM におけるバイアス軽減手法の設計に新たな課題を提示する．

1　はじめに

大規模言語モデル（LLM）の内部表現や出力には，性別や人種に関するバイアスが含まれることが指摘されてきた [1, 2]．近年では，複数の属性が重なり合う交差的バイアスの存在も指摘されている [3]．本研究では基礎的な分析として，性別に基づく個人に対する固定観念や偏見であるジェンダーバイアス [4] に焦点を当てる．加えて，Chen ら [5] は，統計的公平性および因果的公平性の観点からジェンダーバイアス軽減手法を検討した．本研究では，統計的公平性は全ての属性に対して，モデルの出力結果が同等であることを示し，因果的公平性は言語モデルの予測が入力された属性によって左右されないこととする．それらを害するバイアスをそれぞれ統計的バイアス，因果的バイアスとする．これらのバイアスに対し，本研究では LLM のデコーダの出力埋め込みのみへの介入に限定し，それぞれに適したバイアス軽減の操作を設計する．

統計的バイアスは事前学習データの偏りに起因すると考え，本研究では出力埋め込みにおいて確率情報がエンコードされた次元への介入を行う．Cho ら [6] は，出力埋め込みの特定次元に出力確率情報が含まれており，これを除去しても意味的情報が保持されることを示した．この次元を確率操作ベクトルとし，性別を示す対義語間の出力確率が等しくなるよう操作することで統計的バイアスを軽減する．

因果的バイアスは，言語モデル内部でジェンダーの概念を関係ない語彙などと結びつけることに起因すると考え，出力埋め込みからジェンダーの概念を取り除く介入を行う．Park ら [7] は，同一概念を示す対義語の出力埋め込みの差分を概念ベクトルとして捉えることを示した．本研究ではこの考えに基づき，出力埋め込みからジェンダーの概念を表すベクトルを推定し，その成分を除去することで因果的バイアスを軽減する．

本研究では，同一の単語に対して，確率操作ベクトルと概念ベクトルを別個の成分として扱う仮定の下，両バイアス軽減手法の効果を調査する．さらに，言語モデル内部には属性間の結びつきやその方向性が存在することが報告されている [8]．そこで本研究では，提案手法が性別を示す単語に起因するバイアスに加え，性別と関係しない単語に起因するバイアスに対しても有効であるかを検証する．実験の結果，概念ベクトルによるバイアス軽減手法が統計的バイアスに有効であった．特に，性別に関連しない語彙から起因するバイアスに効果的であった．

2　関連研究

LLM に内在するバイアスについては多くの研究が行われ，様々な軽減手法や測定用データセット [9, 10, 11] が提案されてきた．これらは，LLM の学習時，内部表現や出力など，様々な段階への介入を対象としている．例えば，機械翻訳タスクにおいて，エンコーダやデコーダの単語埋め込みから内在的なバイアスを除去しても，下流タスクの公平性が必ずしも改善されないことが示されている [12]．

さらに，LLM の内部表現を用いて，ジェンダーや人種といった高レベルな概念を抽出し，バイアス軽減に応用する研究も行われている．Nguyen ら [13] は，因果モデルに基づいて LLM の内部表現を解釈可能に結びつける手法 [14] を用い，中間層表現から人種概念を表す部分空間を抽出することで，人種に関するバイアスを軽減した．また，Chen ら [15] は，ジェンダーの概念を示す対義語が含まれている文章のペアを LLM に入力し，その際の中間活性の差分を取ることで，ジェンダーの概念を示すベクトルを取り出した．

一方で，既存研究の多くは中間層や内部活性に焦点を当てており，出力確率に直接対応する出力埋め込みや，そこから抽出されるジェンダーの概念を介入対象とする手法は十分に検討されていない．さらに，バイアス研究全体においては，バイアスの定義や対象の曖昧さといった課題も指摘されており [16]，バイアスの種類ごとに手法を設計する枠組みも十分に確立されていない．

3 提案手法

提案手法の概要を図 1 に示す．本手法では，統計的バイアスおよび因果的バイアスに対する介入を，いずれもデコーダの出力埋め込み上で行う．

3.1 統計的バイアス軽減

確率操作ベクトルの算出には，Cho ら [6] が提案した出力埋め込み操作の枠組みを採用した．具体的には，WIKIDPR [17] を用い，平均的なトークン出力確率 $\alpha_{w,\mathcal{D},\theta}$ と出力埋め込み $\mathbf{E}_w^{(o)}$ の間の対数線形関係を重回帰により近似した：

$$-\log \alpha_{w,\mathcal{D},\theta} \approx \mathbf{A}_{\mathcal{D}} \cdot \mathbf{E}_w^{(o)} + \mathbf{B}_{\mathcal{D}} \tag{1}$$

ここで $\mathbf{A}_{\mathcal{D}}$ は回帰係数ベクトルである．この関係に基づき，所望のスケーリング比 r に応じて出力埋め込みを以下のように変換する：

$$\mathbf{E}_w^{(o)\prime} = \mathbf{E}_w^{(o)} - \log(r) \cdot \mathbf{\Omega} \cdot \mathbf{A}_{\mathcal{D}}^{-1} \tag{2}$$

$\mathbf{\Omega}$ は $\mathbf{A}_{\mathcal{D}}$ の各次元の大きさと有意確率に基づく重みベクトルであり，この操作により出力確率を r 倍に調整する．なお，本稿では項 $\mathbf{\Omega} \cdot \mathbf{A}_{\mathcal{D}}^{-1}$ を確率操作ベクトルと呼ぶ．

α_i を女性語群，β_i を男性語群に属する単語（詳細は付録 A.3 参照）とし，ジェンダーの概念を示す n 個の対義語ペアを (α_i, β_i) と定義する．$P(w)$ は語彙 w に対する平均的な出力確率 $\alpha_{w,\mathcal{D},\theta}$ の略記とし，

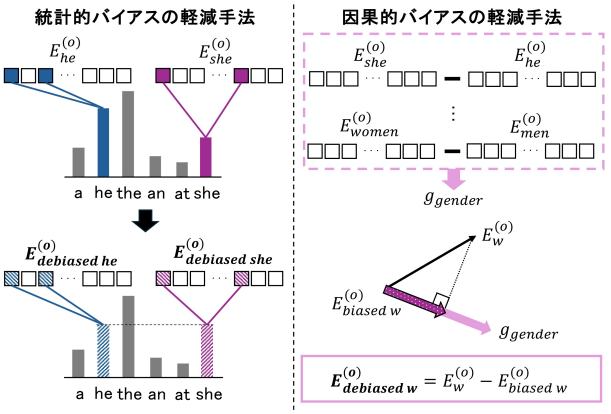

図 1 因果的および統計的バイアスの軽減手法の概要

各ペアのスケーリング係数 $r_{\alpha_i}, r_{\beta_i}$ を以下のように定義する．

まず，各ペアについて $P(\alpha_i), P(\beta_i)$ の平均 $P_{\text{mid},i}$ を求める：

$$P_{\text{mid},i} = \frac{P(\alpha_i) + P(\beta_i)}{2} \tag{3}$$

次に，両者の確率を等しくするためのスケーリング係数 $r_{\alpha_i}, r_{\beta_i}$ を以下のように定める：

$$r_{\alpha_i} = \frac{P_{\text{mid},i}}{P(\alpha_i)}, \qquad r_{\beta_i} = \frac{P_{\text{mid},i}}{P(\beta_i)} \tag{4}$$

これらを前述の変換式に適用することで，各ペアの出力確率を適切に制御する．

3.2 因果的バイアス軽減

概念ベクトルの算出には Park ら [7] の手法を採用した．対義語ペア (α_i, β_i) に対する出力埋め込みをそれぞれ $\mathbf{E}^{(o)}\alpha_i, \mathbf{E}^{(o)}\beta_i$ としたとき，ジェンダーに関する概念ベクトル $\mathbf{g}$ は以下で定義される：

$$\mathbf{g} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{E}_{\alpha_i}^{(o)} - \mathbf{E}_{\beta_i}^{(o)} \right) \tag{5}$$

なお，使用する対義語ペアは，トークナイズ後にトークン ID が一意に定まる単語に限定する．任意の出力埋め込みベクトル $\mathbf{E}_w^{(o)}$ に対し，概念ベクトル $\mathbf{g}$ への正射影ベクトルを計算し，それを元の埋め込みから減算する：

$$\mathbf{E}_{\text{debiased w}}^{(o)} = \mathbf{E}_w^{(o)} - \frac{\mathbf{g} \cdot \mathbf{E}_w^{(o)}}{\|\mathbf{g}\|^2} \mathbf{g} \tag{6}$$

これにより，$\mathbf{E}_w^{(o)}$ の $\mathbf{g}$ 方向の成分が除去される．全語彙に同様の操作を行い，出力空間におけるジェンダー関連表現の影響を低減することを意図する．

確率操作ベクトルと概念ベクトルを個別に適用する場合は，それぞれ「確率操作」と「概念」と表記

表 1　各タスクにおける条件別のテキスト例（共通部分を一部省略）

対象のバイアス	タスク	予測対象	条件とテキスト例
統計的バイアス	タスク 0	性別	**職業明示**: Cherrin began **the attorney's** career [...]. **This person** worked [...].
	タスク 1	性別	**中立**　　: Cherrin began **the person's** career [...]. **This person** worked [...].
因果的バイアス	タスク 2	職業	**性別明示**: Cherrin began **his** career [...]. **He** worked [...].
	タスク 3	職業	**中立**　　: Cherrin began **the person's** career [...]. **This person** worked [...].

する．両方を適用する場合は，順序の影響を検証するため，確率操作ベクトルによる介入後に概念ベクトルによる介入操作を行う「確率→概念」とその逆の「概念→確率」の 2 通りを検証する．

4　実験

4.1　タスク設定

統計的バイアスおよび因果的バイアスの軽減効果を評価するタスク設定を述べる．

タスクに使用するデータ　職業予測のジェンダーバイアスを調査するデータセット Bias in Bios [18] を使用する．職業は全部で 28 種類（表 A.1 参照）あり，本タスクではテスト用のデータを使用する．

タスク設定　統計的バイアスおよび因果的バイアスを評価するため，文中の属性語（性別を示す単語や職業名）が明示されている文と，それらを *the person* に置き換え中立化した文を用意し，以下の 4 つのタスクを設定した（表 1 参照）．**統計的バイアス（性別予測）**では，文中の職業名から性別を予測する設定を**タスク 0**，職業名を中立化した状態で性別を予測する設定を**タスク 1** とする．

一方，**因果的バイアス（職業予測）**では，文中の性別を示す単語から職業を予測する設定を**タスク 2**，性別を示す単語を中立化した状態で職業を予測する設定を**タスク 3** とする．なお，各タスクで使用する詳細な仕様については付録 A.2 に詳述する．

予測結果の算出手法　本研究では，分類ヘッドを付加せず，生成モデルとして用いられる自己回帰型 LLM を対象とする．分類対象の文を LLM に入力した後，*The person is* に続く予測対象語彙の尤度（複数トークンの場合はその和）を算出し，尤度が最大となる語彙をモデルの予測結果とする．

4.2　実験設定

データセット　Bias in Bios のテストデータ 99,000 件のうち，職業ごとに 100 件ずつサンプリングをしたもの（計 2,800 件のデータ）を使用する．

評価指標　提案手法の効果を評価するため，以下の指標を用いる．**統計的バイアス**（タスク 0, 1）では，男女間の尤度差の二乗平均平方根（RMS）によりバイアス強度を定量化する．さらに，米国労働省データ（BLS）[19] とのピアソン相関係数から統計的不平等の解消度を測る．この時，BLS データに対応しない職業については，統計値が取得できないため分析から除外した．なお，本研究では，統計的バイアスは全職業に対する予測確率分布に基づいて評価されるため，性別を示す単語の有無を入れ替えるタスク 2 およびタスク 3 は用いない．

一方，**因果的バイアス**（タスク 2, 3）には正解率（Accuracy）を用いる．これにより，提案手法によるモデル性能の劣化や，性別を示す単語に起因するバイアスの影響を確認する．また，タスク 0 の結果から職業ごとの女性予測確率の平均値を算出し，職業名に起因するバイアスの軽減効果を分析する．

使用モデル　Qwen/Qwen1.5-1.8B-Chat を使用．

表 2　タスク 0 およびタスク 1 の評価結果．

	タスク 0		タスク 1	
手法	RMS(↓)	Pearson(↓)	RMS(↓)	Pearson(↓)
介入なし	0.67	0.61	0.50	0.56
確率操作	0.67	0.61	0.50	0.56
概念	0.48	0.49	0.35	0.45
確率→概念	0.52	0.50	0.38	0.43
概念→確率	0.48	0.49	0.35	0.45

4.3　実験結果

統計的バイアス　タスク 0 とタスク 1 の結果を表 2 に示す．タスク 0 において，提案手法によって RMS もピアソン相関係数も低下したことから，統計的なバイアスを効果的に抑制できたことがわかる．特に概念ベクトルと「概念→確率」の手法が最も効果的であり，次いで「確率→概念」の順となった．確率操作ベクトル単体では効果がないも踏まえると，確率操作ベクトルを先に適用することで，後続の概念ベクトルによるバイアス関連次元の特定が阻害されたためと示唆される．タスク 1 と比較する

表 3　タスク 2 およびタスク 3 の評価結果．Male/Fem. は男女別の正解率，Gap はその差分（Male − Fem.）を示す．

	タスク 2				タスク 3			
手法	Acc	Male	Fem.	Gap(↓)	Acc	Male	Fem.	Gap(↓)
介入なし	64.43	66.49	61.92	-0.0457	65.25	66.23	64.05	-0.0218
確率操作	64.43	66.49	61.92	-0.0457	65.25	66.23	64.05	-0.0218
概念	64.46	66.69	61.76	-0.0493	65.39	66.62	63.90	-0.0272
確率→概念	64.57	66.75	61.92	-0.0483	65.14	66.49	63.50	-0.0299
概念→確率	64.46	66.69	61.76	-0.0493	65.39	66.62	63.90	-0.0272

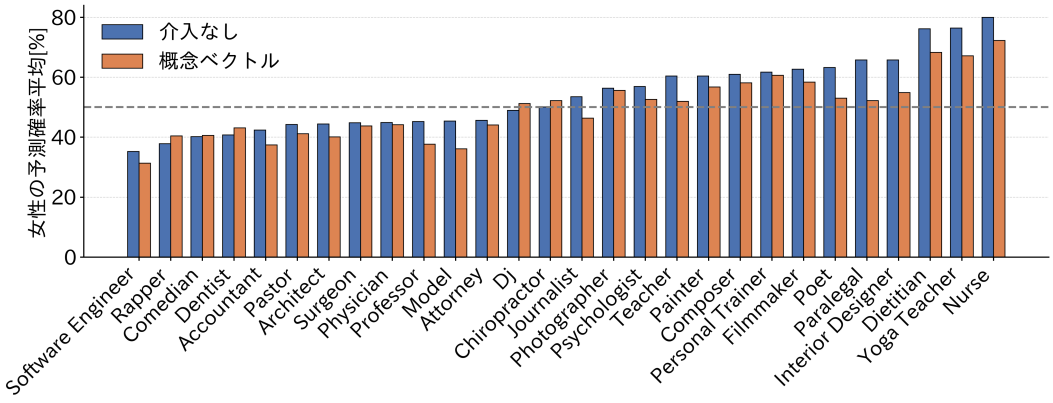

図 2　タスク 0 における職業別女性予測確率の平均値

と，タスク 1 の方が全体的にバイアスが小さく，明示的な職業名をなくすことでバイアス軽減が可能とわかる．その中でも提案手法を用いた際のバイアスが小さいことから，文脈から起因するバイアスに対しても提案手法が効果的であることがわかった．

因果的バイアス　タスク 0 とタスク 1 の結果を表 3 に記す．提案手法を適用しても職業予測の精度は維持され，男女別の精度に関しても同様であったことから，提案手法が LLM の性能自体に与える影響は少ないが，バイアス軽減効果も少ないことがわかる．両タスクを比較すると，タスク 3 の方が全体的な正解率が高く，特に女性の正解率が向上し，男女差も小さくなった．これは性別を示す単語から職業を結びつける因果的バイアスが女性語と強く結びついていることが示唆される．また，性別を示す単語をマスクすることで因果的バイアスを減らすことはできたが，手法間を比較すると，提案手法で同等の効果は得られなかったことがわかる．

図 2 に，バイアス軽減を行わないベースラインと，最も効果の高かった提案手法を用いた際の，女性の予測確率の平均値を示す．女性予測確率が 50% を超える場合，LLM が当該職業を女性と強く関連付けていることを示す．提案手法の適用により，このような職業における女性予測確率が低下し，女性への過度な関連付けが緩和された．以上より，本手法は職業名に起因するバイアスの軽減に有効であることが示唆される．

5　おわりに

本研究では，統計的バイアス軽減のために確率操作ベクトルによる介入を行い，因果的バイアス軽減のために概念ベクトルによる介入を行った．結果として，統計的バイアスには確率操作ベクトルによる介入の効果はなかったが，概念ベクトルを用いた手法は効果的であった．特に，性別を示さない単語（職業名）やその文脈から起因するバイアスに有効だった．反対に，性別を示す単語から起因する因果的バイアスには，どちらの手法も効果がなかった．この要因の一つとして，本研究では概念を線形かつ独立した存在として扱っている点が影響している可能性が考えられる．職業名からジェンダーの概念への対応は比較的一意に定まるのに対し，ジェンダーの概念は他の多様な概念と結びつく可能性があり，その結果として生じる交差的なバイアスを十分に捉えられなかった可能性が示唆される．今後の展望として，より多様な種類のバイアスを対象とし，複雑な概念を扱うために非線形な概念表現を抽出する手法の検討を進める．

謝辞

本研究は，お茶の水女子大学ジェンダード・イノベーション研究所からご支援を頂きました．ここに深謝いたします．

参考文献

[1] Alessandra Miasato and Fabiana Reis Silva. Artificial intelligence as an instrument of discrimination in workforce recruitment. **Acta Universitatis Sapientiae Legal Studies**, 2020.

[2] Shelley J. Correll, Stephen Benard, and In Paik. Getting a job: Is there a motherhood penalty? **American Journal of Sociology**, Vol. 112, No. 5, pp. 1297–1338, 2007.

[3] Lu Cheng, Nayoung Kim, and Huan Liu. Debiasing word embeddings with nonlinear geometry, 2022.

[4] American Psychological Association. Gender bias. https://dictionary.apa.org/gender-bias, 2023. Accessed: 2025-01-01.

[5] Hannah Chen, Yangfeng Ji, and David Evans. Addressing both statistical and causal gender fairness in NLP models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 561–582, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[6] Hakaze Cho, Yoshihiro Sakai, Kenshiro Tanaka, Mariko Kato, and Naoya Inoue. Understanding token probability encoding in output embeddings. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, **Proceedings of the 31st International Conference on Computational Linguistics**, pp. 10618–10633, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

[7] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2024.

[8] Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes, 2024.

[9] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **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. 5356–5371, Online, August 2021. Association for Computational Linguistics.

[10] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models, 2020.

[11] Shaina Raza, Mizanur Rahman, and Michael R. Zhang. Beads: Bias evaluation across domains, 2025.

[12] Bar Iluz, Yanai Elazar, Asaf Yehudai, and Gabriel Stanovsky. Applying intrinsic debiasing on downstream tasks: Challenges and considerations for machine translation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 14914–14921, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[13] Dang Nguyen and Chenhao Tan. On the effectiveness and generalization of race representations for debiasing high-stakes decisions, 2025.

[14] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations, 2024.

[15] Hannah Cyberey, Yangfeng Ji, and David Evans. Unsupervised concept vector extraction for bias control in llms, 2025.

[16] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics.

[17] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau Yih. Dense passage retrieval for open-domain question answering. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics.

[18] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In **Proceedings of the Conference on Fairness, Accountability, and Transparency**, FAT* ' 19, p. 120–128. ACM, January 2019.

[19] U.S. Bureau of Labor Statistics. Labor force statistics from the current population survey, 2024: Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity. https://www.bls.gov/cps/cpsaat11.htm, 2024. Accessed: 2026-01-01.

A　参考情報

A.1　実験に使用した 28 種類の職業

accountant, architect, attorney, chiropractor, comedian, composer, dentist, dietitian, dj, filmmaker, interior designer, journalist, model, nurse, painter, paralegal, pastor, personal trainer, photographer, physician, poet, professor, psychologist, rapper, software engineer, surgeon, teacher, yoga teacher.

A.2　タスク詳細とプロンプト

本研究で設定したタスク 0〜3 の詳細は以下の通りである．

タスク 0 (統計・明示):　文に含まれる職業名などに基づき，文脈から示唆される性別を予測する．実社会における職業別の男女比と比較することで，職業と性別の共起に起因する統計的バイアスがモデルに反映されているかを検証する．

タスク 1 (統計・中立):　本来，文に含まれていた職業名を中立な単語である *person* と入れ替え，その文脈から示唆される性別を予測する．タスク 0 と比較することで，職業名をマスクした場合に統計的バイアスがどの程度変化するかを検証する．

タスク 2 (因果・明示):　文に含まれる性別を示す単語などに基づき，文脈から示唆される職業を予測する．本来，性別情報は職業予測に不要であるため，ここでの予測傾向は性別を示す単語から起因する因果的バイアスを示唆する．

タスク 3 (因果・中立):　本来，文に含まれていた *he/she* を中立な単語である *person* と入れ替え，その文脈から示唆される職業名を予測する．タスク 2 と比較することで，性別情報が職業予測に与える因果的バイアスを検証する．

A.3　概念ベクトル作成用の対義語ペア

表 4 に，実験においてジェンダーの概念を定義するために使用された性別対義語ペアを列挙する．

表 4　ジェンダーの概念を定義するために用いられる対義語ペア．

男性語群	女性語群
actor	actress
batman	batwoman
boar	sow
boy	girl
brother	sister
buck	doe
bull	cow
businessman	businesswoman
chairman	chairwoman
dad	mom
daddy	mommy
duke	duchess
emperor	empress
father	mother
fisherman	fisherwoman
fox	vixen
gentleman	lady
god	goddess
grandfather	grandmother
grandpa	grandma
grandson	granddaughter
groom	bride
he	she
headmaster	headmistress
heir	heiress
hero	heroine
hound	bitch
husband	wife
king	queen
lion	lioness
man	woman
manager	manageress
men	women
mister	miss
murderer	murderess
nephew	niece
poet	poetess
policeman	policewoman
prince	princess
ram	ewe
rooster	hen
sculptor	sculptress
sir	madam
son	daughter
stallion	mare
stepfather	stepmother
superman	superwoman
tiger	tigress
uncle	aunt
valet	maid
waiter	waitress
webmaster	webmistress