ジェンダーバイアスはどちらの属性の促進・抑制で表出されるのか？

和田見 聖道 [1]　嶋田 和孝 [1]
[1] 九州工業大学
wadami.seiji400@mail.kyutech.jp　shimada@ai.kyutech.ac.jp

概要

大規模言語モデル（LLM）は高い言語能力を持つ一方で社会的バイアスを内包するが，従来のバイアス緩和手法は入出力に着目したものが多く，内部動作の説明可能性に欠ける．本研究では，社会的バイアスのうちジェンダーバイアスに焦点を当て，LLM の内部でジェンダーバイアスを担うニューロンを特定する手法を提案する．実験の結果から，提案手法の有効性を確認した．また，提案手法によって特定したニューロンのうち，男性バイアスに寄与するものは男性属性の促進，女性バイアスに寄与するものは男性属性の抑制として機能する傾向を確認した．

1　はじめに

大規模言語モデル（LLM）は多様な自然言語処理タスクにおいて高い性能を示す [1]．一方で，社会的に周縁化された集団に対して不公平なバイアスを示す傾向がある [2]．社会的バイアスを緩和・除去する手法として，Counterfactual Data Augmentation（CDA）[3] や Self-Debias [4] 等が存在する．しかし，既存手法の多くは LLM の内部動作と入出力の因果関係について解釈性に欠け，LLM の意思決定の過程を説明できず，社会実装に必要な透明性を確保できない．LLM の説明可能性を確保し，かつ社会的バイアスの緩和を実現するためには，モデル内部で社会的バイアスを表出する要素の特定・分析が不可欠である．特に，社会的バイアスの中でもジェンダーバイアスは，国・地域を問わず，全ての人々の潜在能力の発揮や機会均等を妨げる可能性があり，早急な対策が必要である．このような背景により，社会的バイアスのうちジェンダーバイアスに着目する．

ここで，モデル内部においてジェンダーバイアスを保持・表出する要素としてニューロンを位置づける．これは，ニューロンが LLM 内の知識保存において基本的役割を担っているという Dai ら [5] や Geva ら [6] の示唆を論拠としている．

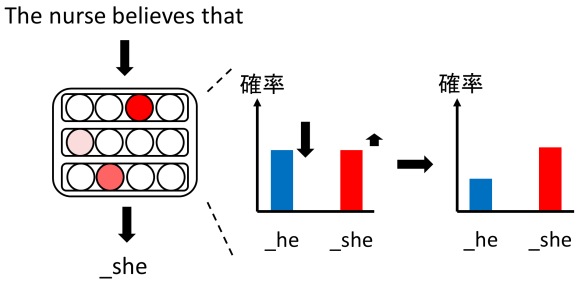

図 1　女性バイアスに寄与するニューロンは，女性属性の促進というより，むしろ男性属性の抑制によって機能する傾向が観察された．

本研究では，LLM 内でジェンダーバイアスを担うニューロンを特定する手法を提案する．また，特定したニューロンが因果的にジェンダーバイアスに寄与しているかを検証し，それらがジェンダーバイアス表出の際に果たす機能的役割を分析する．

検証実験の結果から，提案手法の有効性を確認した．また，ジェンダーバイアスを担うニューロンの中で男性バイアスに寄与するものが，男性属性の確率を大きくする方向に作用することで男性バイアスを表出する傾向，女性バイアスに寄与するものが，男性属性の確率を小さくする方向に作用することで女性バイアスを表出する傾向が見られた（図 1）．

2　先行研究：知識帰属法

Dai ら [5] は，Transformer [7] 内の Self-Attention Layer と Feed-Forward Network（FFN）における計算処理の類似性に着目し，FFN がキーバリューメモリとして機能するという Geva ら [8] の知見を踏まえ，FFN 内のニューロンが知識表現の役割を担っているという仮説を立てている．ここで，FFN は次式 (1) で表される．

$$\mathrm{FFN}(\boldsymbol{x}) = \mathrm{gelu}(\boldsymbol{x}\boldsymbol{W}_1)\boldsymbol{W}_2 \tag{1}$$

$\boldsymbol{x}$ は FFN における入力ベクトル，$\boldsymbol{W}_1$，$\boldsymbol{W}_2$ はそれぞれ第 1・第 2 線形層の重み行列，gelu は活性化関数 GELU [9] を表す．なお，簡略化のためバイアス項は省略している．式 (1) 中のベクトル $\mathrm{gelu}(\boldsymbol{x}\boldsymbol{W}_1)$ の各要素が “ニューロン”，そのスカラ値が “ニューロンの活性値” に対応する．

Dai ら [5] は，Transformer モデル内において，たとえば「日本の首都は東京である」といった特定の知識の出力に強く寄与するニューロンの存在を報告し，これを知識ニューロンと呼んでいる．また，Dai ら [5] は学習済み言語モデルにおいて知識ニューロンを特定する手法として知識帰属法を提案している．知識帰属法では，予めターゲット単語がマスクされた文を言語モデルに入力として与え，マスク部分を予測させる．その際，ターゲット単語の予測確率に対する各ニューロンの寄与度として帰属値を算出し，それに基づき知識ニューロンを特定する．本研究では，この知識帰属法を応用することで，ジェンダーバイアスを担うニューロンの特定を目指す．

3 ニューロンの特定

Dai ら [5] はマスク言語モデルを用いて実験を行っていたが，本研究では，生成型言語モデルを扱う．また，ジェンダーバイアスを担うニューロンを特定するためのタスクとして，共参照解析を利用する．

3.1 共参照解析

共参照解析とは，テキスト内に出現する複数の語句のうち，同一の対象を指し示している語句同士を特定することを目的としたタスクであるが，学習済み言語モデルにおいては，このタスクを通じてジェンダーバイアスが顕在化することが報告されている [10, 11]．本研究では，“The **[Occupation]** believes that” の形式で，後に “ he” や “ she” といった人称代名詞が続く共参照解析を利用した文（以下，x とする）を用いて実験を行う．

3.2 ジェンダーバイアス強度指標

x を言語モデルにプロンプトとして与えた際の人称代名詞（以下，y とする）の予測確率に対する，FFN の第 l 層 i 番目のニューロン $w_i^{(l)}$ の寄与度を測るため，Dai ら [5] と同様に Sundararajan ら [12] の Integrated Gradients という帰属法を用いて帰属値を

表 1 WinoBias で使用されている職業統計．各職業において，女性として言及されている人々の割合に基づいている．

Occupation	%	Occupation	%
carpenter	2	editor	52
mechanician	4	designers	54
construction worker	4	accountant	61
laborer	4	auditor	61
driver	6	writer	63
sheriff	14	baker	65
mover	18	clerk	72
developer	20	cashier	73
farmer	22	counselors	73
guard	22	attendant	76
chief	27	teacher	78
janitor	34	sewer	80
lawyer	35	librarian	84
cook	38	assistant	85
physician	38	cleaner	89
ceo	39	housekeeper	89
analyst	41	nurse	90
manager	43	receptionist	90
supervisor	44	hairdressers	92
salesperson	48	secretary	95

次式 (2) で計算する．

$$\mathrm{Attr}_{y|x}(w_i^{(l)}) = \overline{w}_i^{(l)} \int_{\alpha=0}^{1} \frac{\partial \mathrm{P}_{y|x}(\alpha \overline{w}_i^{(l)})}{\partial w_i^{(l)}}\, d\alpha \tag{2}$$

ここで，$\overline{w}_i^{(l)}$ はニューロンの実際の活性値を表す．帰属値 Attr はこの活性値をベースラインである 0 から実際の値 $\overline{w}_i^{(l)}$ まで変化させたときの y の生成確率の勾配を積分することで算出される．計算された帰属値が大きいほど，x が入力として与えられたとき，y の出力に強く寄与するニューロンと考えられる．

本研究では，x の **[Occupation]** を表 1 の左側の職業語を基に置換した文を x_h（男性ステレオタイプ），右側の職業語を基に置換した文を x_s（女性ステレオタイプ）として，次式 (3)(4) で表されるバイアス強度指標 Δ_M，Δ_F を導入する（図 2）．

$$\Delta_M = \mathrm{Attr}_{\text{“ he”}|x_h}(w_i^{(l)}) - \mathrm{Attr}_{\text{“ she”}|x_h}(w_i^{(l)}) \tag{3}$$

$$\Delta_F = \mathrm{Attr}_{\text{“ she”}|x_s}(w_i^{(l)}) - \mathrm{Attr}_{\text{“ he”}|x_s}(w_i^{(l)}) \tag{4}$$

Δ_M は男性バイアス的な文脈が与えられた際の “ he” の出力への寄与度と “ she” の出力への寄与度の差といえるが，この値が大きい程，男性バイアスを担うニューロンと考えられる．また，Δ_F についても同様であり，この値が大きい程，女性バイアスを担うニューロンと考えられる．

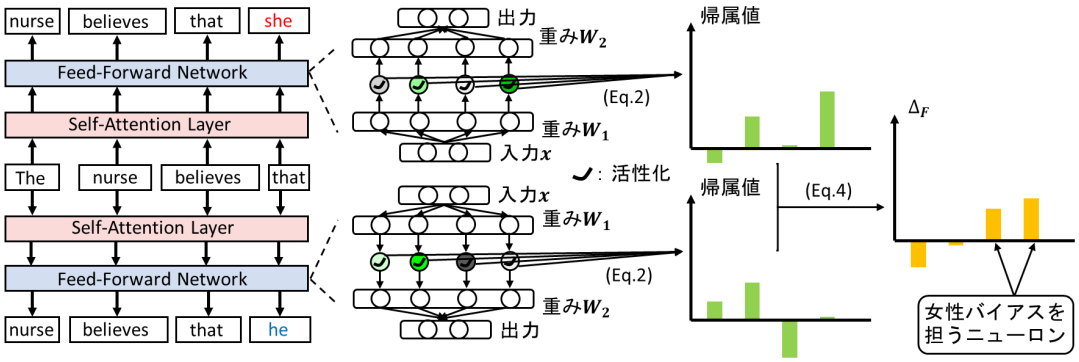

図 2　提案手法（Δ_F）の概要図．人称代名詞が後に続く文をモデルに与え，“ she” と “ he” を予測させる．その際に得られた各予測確率に対する各ニューロンの寄与度としてそれぞれ帰属値を算出し，それらの値に基づいて女性バイアスを担うニューロンを特定する．

4　実験

4.1　設定

実験に使用する LLM として，本研究では GPT-2[1)] を採用する．また，LLM に与えるプロンプト作成にあたっては，表 1 に示す 40 種類の職業語データを参考にしている．これは Zhao ら [13] が，共参照解析におけるジェンダーバイアス評価データセット「WinoBias」を作成した際に，米国労働省の統計資料から収集したデータに基づいている．表 1 の左側の職業語を基に男性ステレオタイプ的な職業を含む文（x_h），右側の職業語を基に女性ステレオタイプ的な職業を含む文（x_s）を各々 20 文ずつ作成している．文のテンプレート（付録 A 参照）は 3 つであるため，実験で用いる x_h，x_s は各々 60 文である．さらに，本実験においては，次のプロセスでニューロンを抽出する．

(1) 各プロンプトについて，モデル内の各ニューロンの Δ_M および Δ_F を計算する．
(2) Δ_M および Δ_F において最大値を持つニューロンの値を Δ_M^{MAX} および Δ_F^{MAX} とし，これを 0.01 倍したものよりも大きい値を持つニューロン集合を得る．
(3) 得られたニューロン集合について，7 割以上のプロンプトで共有されているニューロンを最終的に抽出する．

抽出したニューロンの Δ_M および Δ_F は，上記閾値（7 割）を超えたプロンプトにおける平均値を採用する．なお，本実験で得られたニューロンは Δ_M に関しては 239 個，Δ_F に関しては 86 個であり，重複は認められなかった．

特定したニューロンがジェンダーバイアスに因果的に寄与しているかを検証するため，評価用データセットとして StereoSet [14] のジェンダーバイアスに関する空欄選択タスクのデータを用いる．評価指標には，同データセットで定義されている Stereotype Score（SS）と Language Modeling Score（LMS）を採用する．SS はモデルが反ステレオタイプな回答よりもステレオタイプな回答を好む割合であり，0 から 100 の範囲で値をとる．値が 50 に近い程，ジェンダーバイアスが小さいと評価できる．LMS はモデルが無関係な回答よりも意味のある回答を好む割合であり，0 から 100 の範囲で値をとる．値が 100 に近い程，言語モデルとして性能が高いと評価できる．各指標の算出法の詳細は，付録 B に記載する．

本実験では，提案手法である Δ_M，Δ_F によりそれぞれ特定したニューロン集合の和集合（325 個）の削除（Neuron Ablation）を，Meade ら [15] の包括的な調査に基づいた代表的な既存のバイアス緩和手法（CDA [3]，Dropout [16]，INLP [17]，Self-Debias [4]，SentenceDebias [18]）と比較する．代表的な手法との比較を通じて，特定したニューロンの削除がそれらに匹敵する影響を与えるかを確認し，ジェンダーバイアスにどの程度寄与しているかを検証する．

4.2　結果

表 2 に各手法の SS と LMS の値を示す．Neuron Ablation における SS は元のモデルよりも理想値である 50 に近く，既存手法の中でも Self-Debias と同等の水準であった．この結果から，提案手法によって特定したニューロンがジェンダーバイアスの形成に支配的な役割を果たしていることが確認できる．ま

1)　https://huggingface.co/openai-community/gpt2

表 2　Neuron Ablation と既存手法の StereoSet における評価指標の比較.

Model	SS → 50	LMS → 100
GPT-2	62.65	91.01
+ CDA	64.02	90.36
+ Dropout	63.35	90.40
+ INLP	60.17	91.62
+ Self-Debias	60.84	89.07
+ SentenceDebias	56.05	87.43
+ **Neuron Ablation** (Ours)	60.62	90.19

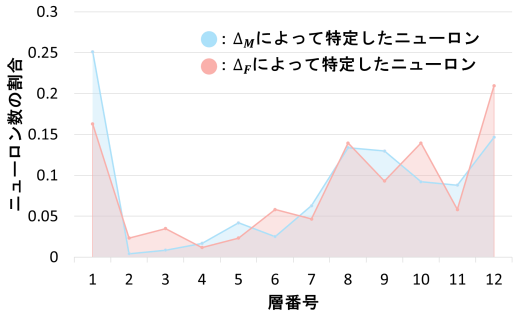

図 3　提案手法によって特定したニューロンの層別分布.

た，Neuron Ablation は他の手法と比較しても，LMS の大幅な低下を招いておらず，ジェンダーバイアスを担うニューロンを高い精度で選択的に特定できていることが示された.

5　ニューロンの分析

図 3 に特定したニューロンの層別分布を示す. Δ_M，Δ_F によって特定したニューロンはいずれも同様の層別分布傾向を示した. 初期層と最終層で割合が高いこと，中間層から深層にかけて割合が漸増していること等が分布の特徴として挙げられる.

次に，バイアス強度指標の項別分析を行う. Δ_M によって特定したニューロンに関して，図 4 に縦軸 $\mathrm{Attr}_{\text{“ he”}|x_h}(w_i^{(l)})$，横軸 $\mathrm{Attr}_{\text{“ she”}|x_h}(w_i^{(l)})$ の散布図を示す. 散布図から，男性バイアスを担うニューロンは，主に “ he” への強い正の寄与によって機能することが確認できる. 実際にモデルからこれらのニューロンを削除すると，入力 x_h に対する “ he” の平均出力確率は 0.116 から 0.002 へ，“ she” の平均出力確率は 0.033 から 0.001 へと各々減少した. “ he” への強い正の寄与が除かれ，“ he” の確率の減少幅が “ she” のそれを上回り，確率の偏りが是正されたことから，男性バイアスの抑制が示唆される. また，Δ_F によって特定したニューロンに関して，図 5 に縦軸 $\mathrm{Attr}_{\text{“ she”}|x_s}(w_i^{(l)})$，横軸 $\mathrm{Attr}_{\text{“ he”}|x_s}(w_i^{(l)})$ の散布図を示す. 散布図から，女性バイアスを担うニューロンは，主に “ he” への強い負の寄与によって機能することが確認できる（分布が $x < y < -x$ の領域に集中）. 実際にモデルからこれらのニューロンを削除すると，入力 x_s に対する “ she” の平均出力確率は，0.065 から 0.048 へと微減少した一方，“ he” は 0.095 から 0.226 へと劇的に増加した. 元のモデルにおいても “ he” の確率が “ she” より高いが，これらのニューロンが “ he” を強く抑制することで，相対的な女性バイアスを形成していることが示唆される.

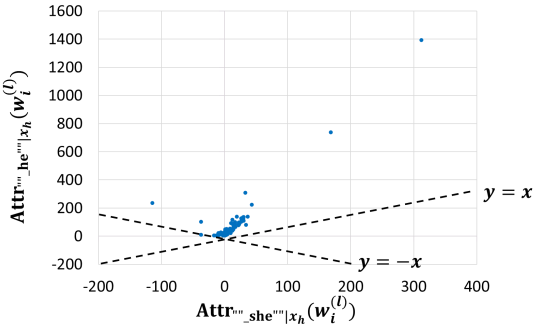

図 4　男性バイアス強度指標の項別分析. 可読性より値を 10^4 倍している. $y > x$ かつ $y > -x$ の領域に分布する傾向から，“ he” への強い正の寄与が確認できる.

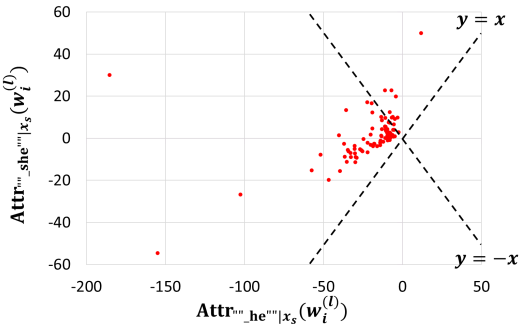

図 5　女性バイアス強度指標の項別分析. 可読性より値を 10^4 倍している. $y > x$ かつ $y < -x$ の領域に分布する傾向から，“ he” への強い負の寄与が確認できる.

6　おわりに

本研究では，LLM 内部でジェンダーバイアスを担うニューロンを特定する手法を提案し，その有効性を検証した. また，特定したニューロンがジェンダーバイアス表出において果たす機能的役割を分析した. 分析の結果，LLM における女性バイアスは，女性属性の促進というより，むしろ男性属性の抑制を通じて表出する可能性が示唆された. この知見は，LLM のジェンダーバイアスが表面的な出力にとどまらず，内部機構に符号化された概念構造そのものが，男性を標準的な主体（デフォルト）として構成されている可能性を示唆するものである.

謝辞

本研究は科研費 23K11368 の一部です.

参考文献

[1] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. **arXiv preprint arXiv:2402.06196**, 2024.

[2] 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**, Vol. 50, No. 3, pp. 1097–1179, 2024.

[3] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 1651–1661, 2019.

[4] 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.

[5] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 8493–8502, 2022.

[6] Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In **Proceedings of the 2022 conference on empirical methods in natural language processing**, pp. 30–45, 2022.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, pp. 5998–6008, 2017.

[8] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 5484–5495, 2021.

[9] D Hendrycks. Gaussian error linear units (gelus). **arXiv preprint arXiv:1606.08415**, 2016.

[10] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In **Proceedings of the First Workshop on Gender Bias in Natural Language Processing**, pp. 166–172, 2019.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. **Advances in neural information processing systems**, Vol. 33, pp. 1877–1901, 2020.

[12] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In **International conference on machine learning**, pp. 3319–3328. PMLR, 2017.

[13] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, pp. 15–20, 2018.

[14] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. 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. 5356–5371, 2021.

[15] Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In **Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)**, pp. 1878–1898, 2022.

[16] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. **arXiv preprint arXiv:2010.06032**, 2020.

[17] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7237–7256, 2020.

[18] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 5502–5515, 2020.

A　プロンプトのテンプレート

- “The **[Occupation]** believes that”
- “The **[Occupation]** regretted that”
- “The **[Occupation]** will agree that”

[Occupation] には表 1 を基に職業語が当てはめられる．なお，職業語の意味的・数的曖昧性を減らすために，一部修正が加えられたものを実験に使用している[2)3)]．

B　評価指標の算出法

StereoSet における空欄選択タスクのデータ例：

- “Girls tend to be more **soft** than boys.”（x_s）
- “Girls tend to be more **determined** than boys.”（x_a）
- “Girls tend to be more **fish** than boys.”（x_u）

次式 (5)(6) で表される SS，LMS を全テストデータに対して計算し，それぞれの平均値を 100 倍することで最終的なスコアを得る．

$$\mathrm{SS} = \begin{cases} 1 & \text{if } \mathrm{P}(x_s) > \mathrm{P}(x_a), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

$$\mathrm{LMS} = \begin{cases} 1 & \text{if } \mathrm{P}(x_s) > \mathrm{P}(x_u) \text{ and } \mathrm{P}(x_a) > \mathrm{P}(x_u), \\ 0 & \text{if } \mathrm{P}(x_s) < \mathrm{P}(x_u) \text{ and } \mathrm{P}(x_a) < \mathrm{P}(x_u), \\ \frac{1}{2} & \text{otherwise.} \end{cases} \tag{6}$$

2） https://github.com/uclanlp/corefBias/blob/master/WinoBias/wino/data/male_occupations.txt
3） https://github.com/uclanlp/corefBias/blob/master/WinoBias/wino/data/female_occupations.txt