文脈内学習に基づく大規模言語モデルの性別バイアス抑制

大葉大輔 [1]　金子正弘 [2]　Danushka Bollegala[3]
[1] 東京大学生産技術研究所　[2]MBZUAI　[3]University of Liverpool
oba@tkl.iis.u-tokyo.ac.jp　Masahiro.Kaneko@mbzuai.ac.ae
danushka@liverpool.ac.uk

概要

大規模言語モデル (LLM) は懸念されるレベルの性別バイアスを内包している。先行研究では LLM の追加学習や復号化器の改変に基づくバイアス除去手法が提案されているが、GPT-4 のような非公開の LLM の場合、内部のパラメータやモジュールを利用できない。本稿では、テンプレートと実世界の統計情報を用いて構築したプリアンブルを LLM の入力に追加するだけで、性別的に偏見のある生成を防ぐ手法を提案する。提案プリアンブルは、統計において特定の性別に偏りがある対象を反実仮想的または中立的に記述する。英語 LLM を用いた評価では、下流タスクの性能への悪影響を抑えながら、性別バイアスを含んだ生成を抑制できることを示した。

1　はじめに

大規模言語モデル (LLM) は深刻なレベルの社会的バイアスを含むことが報告されている [1, 2, 3]。これまでに数多くのバイアス除去手法が提案されてきた; パラメータの追加学習 [4, 5, 6, 7]、ランダムノイズの適用 [8]、有害単語の生成確率の調整 [2]、学習データの拡張 [9, 10, 11]。一方で、セキュリティや商業的利益の観点から全ての LLM がパラメータやモジュールを提供しているわけではなく (e.g., GPT-4)、生成確率を調整するための復号化プロセスの修正 [2] も行えないことがある。その場合、非公開 LLM のエンドユーザーが出来ることは、特定したバイアスをモデルの所有者に報告し、修正を待つことのみである。また、たとえバイアス除去のためにパラメータを追加学習できたとしても、下流タスクの性能低下や異なる社会的バイアスの増幅などの予期せぬ副作用が生じる可能性がある。

本稿では、テキストプリアンブルを LLM の入力先頭に追加するだけで、特定の社会的バイアスを含んだ生成を防ぐ手法を提案する（図 1、節 2)。これは、LLM のパラメータや復号化器を利用する必要がないだけでなく、エンドユーザーが独立に使用することができる。我々は、社会的バイアスの実例として “性別バイアス” に注目し、反実仮想型と説明文型の 2 種類のプリアンブルを導入する。反実仮想型プリアンブルは、LLM の認識を反偏見的な方向に修正するために、実世界で偏見的性別連想があるとされる対象を反実仮想的に記述する。説明文型プリアンブルは、偏見的性別連想のある対象が本来性別に依存しないことを LLM に伝えるため、それら対象を性別非依存な単語を用いて中立的に記述する。より具体的には、米国市民人口統計を用いて偏見的性別連想がある対象（e.g., 職業）を収集し、複数のテンプレートを用いてプリアンブルを構築した。

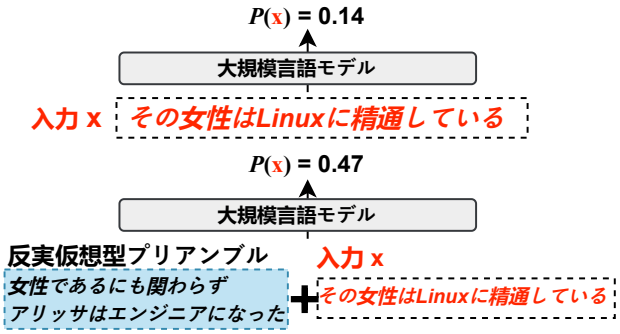

図 1: 提案手法適用時 (下)、非適用時 (上) の概念図。反偏見的な文の尤度が上がることを期待する。

実験では、複数の英語 LLM (MPT [12]、 OpenLLaMA [13]、LLaMA2 [14]) に提案手法を適用し、CrowsPairs ベンチマーク [15] を用いてバイアス抑制効果を検証した (節 4)。その結果、下流タスク (COPA [16]、HellaSwag [17]) の性能劣化を抑えながら、特定の性別に対する偏見的な応答生成を抑制できることが示された。さらに、バイアスを抑えるように直接的に命令する指示文よりも提案手法が効果的であることや、提案手法の効果を最大化する上で LLM の賢さが重要な要素であることを示した。

表 1: 職業人口分布が女性に偏っている “*childcare workers*” を対象に構築したプリアンブルの例。

反実仮想型-simple	*Donald became a childcare worker.*
反実仮想型-detailed	*Despite being a male, Donald became childcare worker.*
説明文型-simple	*Childcare workers look after children.*
説明文型-detailed	*Childcare workers participate in supervising children to provide safety .*

2 文脈内学習に基づくバイアス抑制

プリアンブルを入力先頭に追加することで、追加学習や復号化器の修正を行うことなく性別バイアスの含んだ生成を防ぐ手法を提案する。その過程で、**反実仮想型**と**説明文型**のプリアンブルを導入する。

反実仮想型: LLM の認識を反偏見的な方向に歪めることを意図して、実世界の偏見的性別連想と矛盾するプリアンブルを導入する (表 1; 上部)。偏見的性別連想として、性別分布が偏った “職業” を利用する。プリアンブル作成には詳細度の異なる以下のテンプレートを用いる:

反実仮想型-simple

i): {M-NAME} *became a(n)* {F-JOB}.

ii): {F-NAME} *became a(n)* {M-JOB}.

反実仮想型-detailed

iii): *Despite being a male*, i)

iv): *Despite being a female*, ii)

ここで、M-/F-NAME は男性または女性に多く付けられた名前、M-/F-JOB は職業人口分布が男性または女性に偏っている職業を指す。これらは米国市民を対象とした統計情報[1)2)] を基に抽出する（付録 A)。

説明文型: 職業などの偏見的性別連想のある対象が、本来性別に中立であることを LLM に伝えるため、性別依存な単語 (e.g., *he*) を使わずにそれら対象を記述するプリアンブルを導入する (表 1; 下部)。ここでも、分布が性別的に偏った “職業” に着目し、詳細度の異なる 2 通りの説明文を人手で作成する; 職業名+3 単語 (simple)、+7 単語 (detailed)。

補足: 本論文では、身体的なものを含む様々な観点の性別バイアスの存在を認めつつ、利用容易性から “職業に紐づく性別バイアスデータ” を使用した。実世界統計の詳細を付録 A に、またプリアンブルの実例を（表 1 に加え）付録 B に記載する。

1) https://www.bls.gov/cps/cpsaat11.htm
2) https://namecensus.com

3 生成型 LLM のバイアス評価尺度

これまで、性別バイアスの評価尺度はマスク言語モデル (MLM) を対象に設計されてきた [18, 15, 19]。これらは、我々が扱う生成型 LLM (e.g., LLaMA2) に直接適用することが困難である。本節では、既存の評価尺度を生成型 LLM のために調整した “絶対バイアススコア” を導入する。また、バイアス抑制効果をより敏感に測定することのできる “相対バイアススコア” を併せて導入する。

前提: バイアス評価データ D に含まれる文ペアを (s, a)、偏見的な文を s、反偏見的な文を a とする。[3)] また、提案プリアンブルの使用・不使用を表す条件フラグを cc および nc とし、LLM のパラメータを θ とする。そして、cc または nc という条件下で、LLM が文 s に対して算出する尤度をそれぞれ $P(s|\theta, cc)$ および $P(s|\theta, nc)$ と表記する。

絶対バイアススコア: MLM のための評価尺度を用いた文脈では、a の尤度よりも s の尤度が大きい文ペアの割合を評価することが一般的である。本論文の文脈では cc または nc 条件によってこの割合がどう変化するかを観測すればよい。ただし、尤度の計算方法は MLM と生成型 LLM では異なる。ここでは、Teacher-forcing [20] に基づいて s および a の尤度計算を行う。計算された尤度を基に、絶対スコアは以下式で算出される:

$$\text{絶対スコア}_{nc} = \frac{1}{|D|} \sum_{(s,a)} \mathbb{I}[P(s|\theta, nc) \geq P(a|\theta, nc)] \quad (1)$$

$$\text{絶対スコア}_{cc} = \frac{1}{|D|} \sum_{(s,a)} \mathbb{I}[P(s|\theta, cc) \geq P(a|\theta, cc)] \quad (2)$$

$\mathbb{I}[x]$ は x が True なら 1 を、False なら 0 を返す。

相対バイアススコア: 絶対バイアススコアでは、s の尤度と a の尤度の大小関係を覆さないレベルのバイアス抑制効果には鈍感である。[4)] そこで、尤度の大小関係を連続的に表現したスコアを導入する:

$$\text{相対スコア}_{nc} = \frac{1}{|D|} \sum_{(s,a)} \log \frac{P(s|\theta, nc)}{P(a|\theta, nc)} \quad (3)$$

$$\text{相対スコア}_{cc} = \frac{1}{|D|} \sum_{(s,a)} \log \frac{P(s|\theta, cc)}{P(a|\theta, cc)} \quad (4)$$

実験（節 4）では、両スコアから得られる傾向が大きくは異ならないことを確認した。

3) 例: s = “彼は医者だ”, a = “彼女は医者だ”

4) 例: $P(s|\theta, nc) = 0.63$、$P(a|\theta, nc) = 0.21$、$P(s|\theta, cc) = 0.48$、$P(a|\theta, cc) = 0.41$。この場合、バイアス抑制効果は絶対スコアには反映されない。

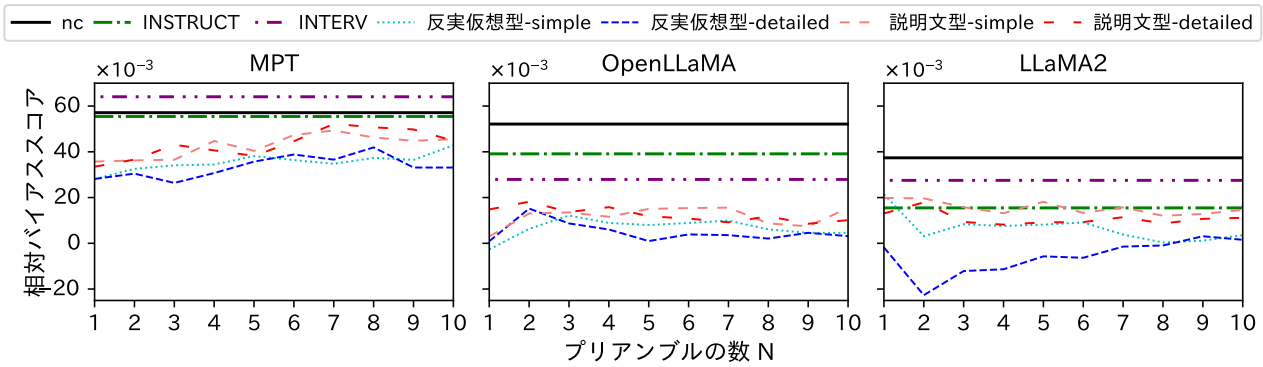

図 2: 相対バイアススコア。N 個のプリアンブルは事前計算した perplexity に基づいて選択・順序付けした。

表 2: MMLU [21]、TruthfulQA; TQA [22]、 ARC [23]、HellaSwag; HS [17] におけるベンチマーク結果。

Model	Avg.↑	MMLU↑	TQA↑	ARC↑	HS↑
MPT-7B	47.4	30.8	33.4	47.7	77.6
OpenLLaMA-7B	48.2	41.3	35.5	43.7	72.2
LLaMA2-7B	**54.3**	**46.9**	**38.8**	**53.1**	**78.6**

4 実験

提案手法を複数の英語 LLM に適用しバイアス抑制効果を検証する（節 4.2）。また、下流タスクに与える影響を性能劣化の観点から検証する（節 4.3）。

4.1 設定

LLM: 形態素の複雑さが限定的な英語に焦点を当て、英語の LLM を評価に使用した。具体的には、基本性能（表 2）が異なる 3 つのモデルを採用した; MPT-7B [12]、OpenLLaMA-7B [13]、LLaMA2-7B [14]。表 2 の値は Open LLM Leaderboard[5] から引用した。

バイアス抑制の評価データ: 様々な社会的バイアスを対象とした偏見的な文 (s) と反偏見的な文 (a) のペアから構成される CrowsPairs データセット [15] を利用した。具体的には、性別バイアスに関係する 262 件のペアを使用した (i.e., $|D| = 262$)。

下流タスクの評価データと尺度: 常識推論・因果関係に関係するベンチマーク; COPA [16] および HellaSwag [17] を使用した。MosaicML[6] に従って分類器を別途訓練せずにゼロショット評価を行い、提案手法の適用による性能の変化 (ΔAcc.) を報告する。

プリアンブルの設定: 各プリアンブルタイプ (反実仮想型-*/説明文型-*) ごとに N 個のプリアンブルを連結して使用した。ここで、効果的な N 個のプリアンブルの選び方および並べ方は自明ではない。そこで、事前実験の結果から、各タイプごとに最も低い perplexity[7] を持つ N 個を選び、左から昇順で連結した。事前実験ではランダム選択の場合よりも優秀なバイアス抑制性能を達成できた（付録 D）。

比較手法: **INSTRUCT** [24] は入力先頭に指示文を与える; "*Please ensure that the following is not biased and does not involve gender stereotyping*"。**INTERV** [25] は指示文に加え、BBQ [26] から抽出した偏見的な質問応答デモと反偏見的なデモを 4 つずつ提示する。

4.2 結果: バイアス抑制の効果

図 2 に相対バイアススコアの傾向を示す。全ての LLM において、提案プリアンブルの利用により *nc* の場合よりも性別バイアスが減少した。紙面の都合上、絶対スコアの傾向については付録 C に回すが、以下で述べる傾向と概ね同様のものが観測された。

反実仮想型 vs. 説明文型: 反実仮想型プリアンブルはより優れたバイアス抑制効果を示した。これは、偏見的性別連想に中立的な記述よりも、それに逆行する記述の方が、バイアス抑制のための情報を多く含むことを示唆する。また、金子ら [27] は WordNet [28] から抽出した職業定義文を用いて単語埋め込みのバイアスを除去した。本実験結果は、単語埋め込みの優れたバイアス除去性能は反実仮想的な例を用いても達成できることを示唆する。

Simple vs. Detailed: LLaMA2 において顕著な傾向として、詳細なプリアンブル (*-detailed) を用いた方がより多くのバイアスを抑制できることが示

5) https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
6) https://github.com/mosaicml/llm-foundry/tree/main/scripts/eval
7) プリアンブル選定のための perplexity 計算は、本評価のインスタンス (e.g., s や a) とは独立かつ事前に実行した。

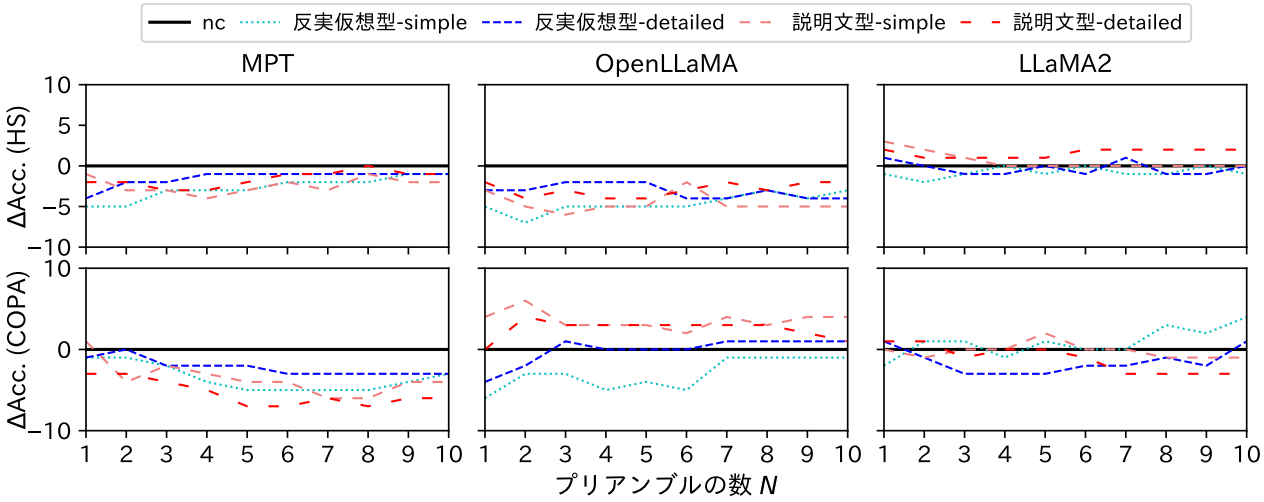

図 3: 提案手法適用時の下流タスク（COPA・HellaSwag; HS) の性能劣化。ブリアンブル構成は図 2 と同じ。

された。これは、入力系列長が長くなることで計算コストが増加するというトレードオフはあるものの、詳細な記述はより正確にプリアンブルの意図を LLM に伝えることができることを示唆している。

N の変動: プリアンブルの数 N を変化させた時、$N \leq 3$ の時に最もバイアスを抑制できた。ランダム選択の結果（付録 D）との比較からも、単純なヒューリスティックス (i.e., perplexity) を用いることで効果的なプリアンブルを選定できることを示している。一方で、N の増加に対する性別バイアスの単調減少は見られなかった。これはプリアンブルの冗長性を考慮することへの課題を示している。

LLM 間の比較: 提案手法によって、LLaMA2 が、その次に OpenLLaMA が最も小さな性別バイアスを達成した。これは、賢い LLM ほど（表 2）プリアンブルの意図を理解できるためと考えられる。また、こうした LLaMA2 や OpenLLaMA は MPT に比べて生来の (*nc*) バイアスが小さいことも示された。

対比較手法: **INSTRUCT** は提案手法に劣った。これは、事前学習済み言語モデルは “否定” の理解が難しいという Kassner ら [29] の報告に理由を見出すことができる。LLaMA2≥OpenLLaMA≥MPT の順で INSTRUCT の効果が見られたのは、より賢い LLM はより良い指示追従能力を持つためと考えられる。**INTERV** も提案手法に劣った。反偏見的なデモと偏見的なデモの両方を均等に提示する INTERV は、反実仮想的な記述に比べて LLM に驚きがないのかもしれない。また、INTERV の方針は、性別的に中立な意図を伝える説明文型と通ずるものがあるが、説明文の方が有用な資源であることを示唆している。

4.3 結果: 下流タスク性能への影響

図 3 から、提案手法は性別バイアスを含んだ生成を防ぎながらも (節 4.2)、下流タスクの性能劣化を最良で 0%、最悪でも-7%に留めることが分かる。

異なるプリアンブルタイプ間の比較では、バイアス抑制評価において良好な結果を示していた反実仮想型-detailed が最も下流タスクの性能を低下させなかった（全ての LLM およびデータに渡って最悪でも-4%の劣化)。異なる LLM 間での比較では、LLaMA2 が最も性能劣化を抑えていた。これは、節 4 の結果も考慮すると、“バイアス抑制効果を高めること”、および “下流タスクへの影響を抑えること” の両観点において LLM の基本性能（表 2）が重要な要素であることを示唆している。

5 おわりに

本稿ではプリアンブルを LLM の入力先頭に追加することで、性別的偏見を持ったテキスト生成を防ぐ手法を提案した。プリアンブルは、実世界統計とテンプレートを用いて構築され、反実仮想型および説明文型の 2 タイプからなる。CrowsPairs を用いた実験では、提案手法が英語 LLM (e.g., LLaMA2) の性別バイアスを抑制できることを示した。また、下流タスク (e.g., COPA) の性能に与える負の影響が限定的であることを示した。分析では、提案プリアンブルは指示文よりも有効なことや、LLM の基本性能が提案手法の効果を引き出すことを示した。

社会的バイアスには言語特有のものがある [30]。今後はプリアンブルの多言語化に取り組みたい。

謝辞

本研究は JSPS 科研費 22KJ0950 の助成を受けたものです。

参考文献

[1] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In **EMNLP-IJCNLP**, pp. 3407–3412, 2019.

[2] Timo Schick, Sahana Udupa, and Hinrich Schütze. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. **TACL**, Vol. 9, pp. 1408–1424, 2021.

[3] Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In **NAACL**, pp. 609–614, 2019.

[4] Masahiro Kaneko and Danushka Bollegala. Debiasing pre-trained contextualised embeddings. In **EMNLP**, pp. 1256–1266, 2021.

[5] Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In **Findings of ACL-IJCNLP**, pp. 4534–4545, 2021.

[6] Anne Lauscher, Tobias Lueken, and Goran Glavaš. Sustainable modular debiasing of language models. In **Findings of EMNLP**, pp. 4782–4797, 2021.

[7] Yue Guo, Yi Yang, and Ahmed Abbasi. Auto-debias: Debiasing masked language models with automated biased prompts. In **ACL**, pp. 1012–1023, 2022.

[8] 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:2010.06032**, 2020.

[9] Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In **ACL**, pp. 1651–1661, 2019.

[10] Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In **EMNLP-IJCNLP**, pp. 5267–5275, 2019.

[11] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In **NAACL**, pp. 629–634, 2019.

[12] MosaicML NLP Team, et al. Introducing mpt-7b: A new standard for open-source, ly usable llms, 2023.

[13] Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama, May 2023.

[14] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. **arxiv:2307.09288**, 2023.

[15] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In **EMNLP**, pp. 1953–1967, 2020.

[16] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **AAAI spring symposium**, pp. 90–95, 2011.

[17] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In **ACL**, pp. 4791–4800, 2019.

[18] Masahiro Kaneko and Danushka Bollegala. Unmasking the mask–evaluating social biases in masked language models. In **AAAI**, Vol. 36, pp. 11954–11962, 2022.

[19] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In **ACL-IJCNLP**, pp. 5356–5371, 2021.

[20] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. **Neural computation**, Vol. 1, No. 2, pp. 270–280, 1989.

[21] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. **arxiv:2009.03300**, 2020.

[22] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In **ACL**, pp. 3214–3252, 2022.

[23] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. **arxiv:1803.05457**, 2018.

[24] Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas 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:2302.07459**, 2023.

[25] Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. Prompting gpt-3 to be reliable. **arxiv:2210.09150**, 2022.

[26] 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 **Findings of ACL**, pp. 2086–2105, 2022.

[27] Masahiro Kaneko and Danushka Bollegala. Dictionary-based debiasing of pre-trained word embeddings. In **EACL**, pp. 212–223, 2021.

[28] Christiane Fellbaum. Wordnet. In **Theory and applications of ontology**, pp. 231–243. Springer, 2010.

[29] Nora Kassner and Hinrich Schütze. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In **ACL**, pp. 7811–7818, 2020.

[30] Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. Gender bias in masked language models for multiple languages. In **NAACL**, pp. 2740–2750, 2022.

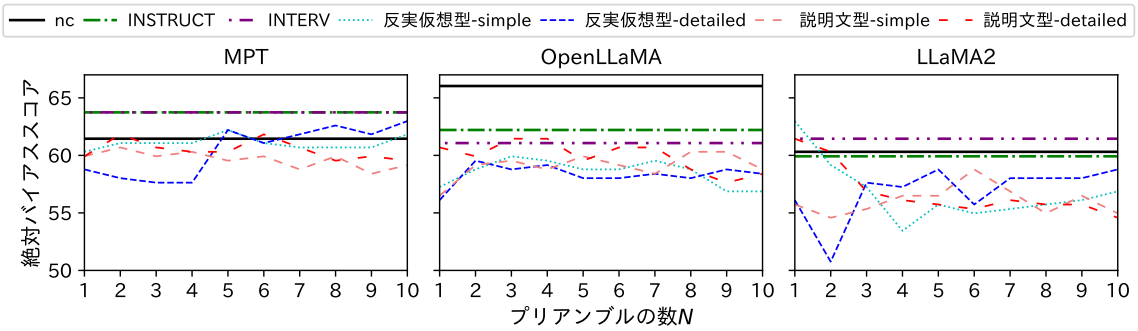

図 4: 絶対バイアススコア。N 個のプリアンブルは事前計算した perplexity に基づいて選択・順序付けした。

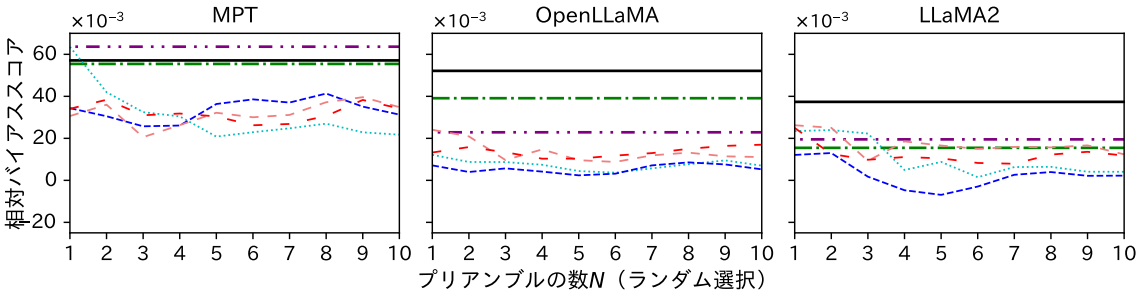

図 5: 相対バイアススコア。N 個のプリアンブルはランダムに選択した。

表 3: 統計に性別的偏りのある名前と職業の一部。

M-NAME:	*Noah, Donald, Eric, Joshua, Kyle, Jordan, ...*
F-NAME:	*Lauren, Lisa, Victoria, Karen, Dawn, Jasmine, ...*
M-JOB:	*Firefighter, Police officer, Pest control worker, ...*
F-JOB:	*Human resources manager, Medical assistant, ...*

表 4: 構築したプリアンブルの実例。

反実仮想型-simple	*John became a teaching assistant.*
反実仮想型-detailed	*Despite being a male, John became a teaching assistant.*
説明文型-simple	*Teaching assistants facilitate student learning.*
説明文型-detailed	*Teaching assistants provide support in education to facilitate learning.*

A　使用した統計データ

我々は米国労働統計局が収集する人口労働統計[8)]から、男性が 70 %以上、または女性が 70 %以上の労働者を占める職業を男性的職業; M-JOB、女性職業; F-JOB として抽出した。また、米国国勢調査局が提供する人口統計情報[9)]に基づいて、1970 年から 2000 年に生まれた子供につけられた名前のうち、頻出上位 30 位までを F-NAME または M-NAME として抽出した。表 3 に抽出した職業/名前の一部を示す。

B　プリアンブルの構築と実例

性別的に偏った職業と名前 (付録 A) を基に、第 2 節に記載の方法により、反実仮想型ならびに説明文型プリアンブルを構築した。本文の表 1 に加え、構築したプリアンブルの一部を表 4 に例示した。

C　絶対バイアススコアの傾向

絶対バイアススコアに基づくバイアス抑制効果の結果を図 4 に示す。概して、i) 提案手法のベースラインに対する優位性、ii) 異なる提案プリアンブルタイプ間での優位性、iii) プリアンブル数 N に対するスコアの変化など、本文記載の相対スコアで得られたもの（図 2）と同様の傾向であった。

D　プリアンブルの選択方法の比較

各プリアンブルタイプごとに、構築済みののプリアンブル集合から N 個のプリアンブルをランダムに選択し、使用した。相対バイアススコアについて、3 回試行平均の結果を図 5 に示す。本文記載の、事前に計算した perplexity の値を使った選択・並び替え（図 2）に比べ、N が小さい時のバイアス抑制効果が弱いことが示された。

8) https://www.bls.gov/cps/cpsaat11.htm
9) https://namecensus.com/