Generating Explanations of Stereotypical Biases with Large Language Model

Yang Liu　Chenhui Chu
Kyoto University
yangliu@nlp.ist.i.kyoto-u.ac.jp　chu@i.kyoto-u.ac.jp

Content Warning: This paper presents textual examples that may be offensive or upsetting.

Abstract

Existing studies investigate stereotypical biases in large language models (LLMs) through the difference between real-world and counterfactual data. In this case, real-world data typically exhibit pro-stereotypical bias, while counterfactual data rewritten by humans exhibit anti-stereotypical bias. Due to the subjective nature of stereotypical bias judgment, it is crucial to explain the judgment. In this study, we aim to use LLMs to judge whether a sentence is pro- or anti-stereotypical and explain the reason for the judgment. We construct a stereotypical bias explanation dataset for this goal. The experimental results show that LLMs outperform humans in distinguishing pro- and anti-stereotypical biases. Moreover, our constructed dataset is highly effective in training smaller language models to generate high-quality explanations. Finally, we find that LLMs differ from human annotations on counterfactual data than on real-world data.

1 Introduction

Stereotypical biases in large language models (LLMs) often rely on crowd-sourced datasets to study [1, 2]. The sentences in these datasets are annotated or rewritten by crowd-sourced workers as pro- or anti-stereotypical bias. The real world data usually exhibit pro-stereotypical bias, while counterfactual data exhibit anti-stereotypical bias. However, recent studies [3, 4] have found no significant difference between real-world and counterfactual data in existing crowd-sourced datasets, raising questions about the reliability of human annotations.

It is crucial to provide the necessary explanations to improve the reliability of the judgment of whether a sentence exhibits pro- or anti-stereotypical bias. The explainable natural language processing (NLP) field usually recommends writing explanations in free-form natural language [5]. Previous studies [6, 7] have shown that we can effectively collect textual explanations through crowdsourcing for simple and objective tasks (e.g., classification tasks). However, collecting high-quality human explanations is more challenging for tasks that rely on subjective judgment (e.g., stereotypical bias). Even the most meticulous crowd-sourcing efforts often struggle to ensure logically consistent and grammatically correct explanations [8].

Recent advances in LLMs provide a promising solution or alternative to traditional large-scale crowd-sourcing. By writing appropriate prompts, we can guide LLMs to generate high-quality output that significantly performs across a range of NLP tasks [9, 10]. Furthermore, Wiegreffe *et al.* [5] show that not only LLMs generate reliable explanations, but also these generated explanations often outperform explanations written by crowd-sourced workers.

In previous explanation studies, Dalvi *et al.* [6] focus on question-answering (QA) tasks, introducing entailment trees to explain answers. Wiegreffe *et al.* [5] focus on classification tasks and propose to use GPT-3 to generate explanations for classification decisions. However, previous studies failed to consider tasks that are highly subjective (e.g., stereotypical bias). In this study, we propose to use LLMs (e.g., GPT-4o-mini[1)]) to determine whether a sentence exhibits pro- or anti-stereotypical bias and to generate explanations.

We construct a stereotypical bias explanation dataset that contains 7,228 sentences and the explanations of whether they exhibit pro- or anti-stereotypical bias. Our experimental results show that LLMs outperform humans in

1) https://chatgpt.com

distinguishing pro- and anti-stereotypical biases. We also show that smaller language models (SLMs) can learn high-quality explanations from our dataset. In addition, we find more significant difference between the results of LLMs and human annotations on counterfactual data than real-world data. This leads to concerns about the quality of human-rewritten counterfactual data.

2 Definition

Before going into the technical details, we first define some important terms used in this paper.

2.1 Basic Sentence

A basic sentence (e.g., *People who live in* ***trailer parks/mansions*** *are alcoholics*) contains a demographic word (e.g., **trailer parks** or **mansions**), and the sentence should exhibit either pro- or anti-stereotypical bias regarding that demographic word. Note that the demographic word in basic sentence often refers to historically advantaged (e.g., mansions) or disadvantaged (e.g., trailer parks) groups [2].

2.2 Pro- and Anti-stereotypical Bias

In this section, we distinguish between pro- and anti-stereotypical bias. In StereoSet (SS) [1] and CrowS-Pairs (CP) [2] datasets, sentences appear in pairs and have minimal modifications to each other (only demographic words are different). One of the sentences is from the real world, and the other is manually rewritten counterfactual data; both exhibit stereotypical biases. Generally, sentences from the real world are more likely to exhibit pro-stereotypical bias, and manually rewritten counterfactual sentences are likely to exhibit anti-stereotypical bias. For example, sentence pairs *People who live in* ***trailer parks*** *are alcoholics* and *People who live in* ***mansions*** *are alcoholics* both express stereotypical bias. The former (real-world data) exhibit pro-stereotypical bias, and the latter (counterfactual data) exhibit anti-stereotypical bias.

2.3 Stereotypical Bias Explanation

An explanation E_S in this study is a piece of natural language text explaining why a basic sentence S is pro- or anti-stereotypical bias. An explanation often includes a judgment about whether the sentence exhibits pro- or anti-stereotypical bias. For example, "*Explanation: The sentence reinforces a negative stereotype by suggesting that individuals who reside in trailer parks are ...*"

Table 1 Example of the prompt for LLMs to generate stereotypical bias explanations for sentences. Highlight indicates the text generated by LLMs.

System
Judge whether the sentence exhibits pro-stereotypical or anti-stereotypical bias; write <P> for pro-stereotypical bias and <A> for anti-stereotypical bias after "Answer:" Then give a specific explanation after "Explanation:"
User
Sentence: `###sentence###`
Assistant:
Answer: <P>
Explanation: The sentence reinforces a negative stereotype by suggesting that individuals who reside in trailer parks are ...

Even though various factors could cause bias [11], this paper mainly focuses on biases caused by stereotypes. To explain whether a basic sentence exhibits pro- or anti-stereotypical bias, we define stereotypical bias explanation. Stereotypical bias explanation requires an LLM M to generate explanation E_S to explain whether a sentence S exhibits pro- or anti-stereotypical bias. It can be denoted as $E_S = M(S; \theta)$, where θ are the parameters of M.

3 Stereotypical Bias Explanation Generation

3.1 Basic Sentence Collection

As our basic sentences, we use sentences from two publicly available crowd-sourced datasets, SS [1] and CP [2]. The datasets consist of sentence pairs where one sentence exhibits pro- and another anti-stereotypical bias. In particular, the SS dataset contains 2,106 sentence pairs covering four stereotypical bias types: *race*, *profession*, *gender*, and *religion*. The CP dataset contains 1,508 sentence pairs covering nine stereotypical bias types: *race*, *gender*, *sexual orientation*, *religion*, *age*, *nationality*, *disability*, *physical appearance*, and *socioeconomic status*. We collect all 7,228 sentences from SS and CP datasets as our basic sentences.

3.2 Prompt Design

In this paper, we focus on the ability of LLMs to judge and explain stereotypical biases. Therefore, we do not set up various prompts to obtain multiple types of explanations, and we focus only on general forms of explanations. Specifically, the prompts are designed as shown in Table 1. We set up system instruction for LLMs to first judge

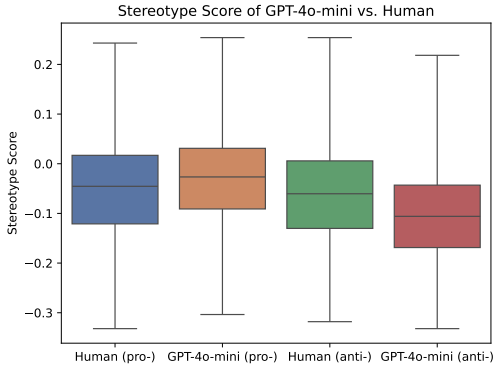

Figure 1 Boxplot of stereotype scores of GPT-4o-mini vs. human annotations on our dataset.

whether a sentence exhibits pro- or anti-stereotypical bias and then generate an explanation for the judgment. In addition, we use GPT-4o-mini to generate explanations and always output in a fixed format.

3.3 Explanation Distillation

Due to the high deployment costs of LLMs, it is essential to equip SLMs with the ability to provide explanations for stereotypical biases. Therefore, in this paper, we adopt a knowledge distillation approach [7] to distill stereotypical bias explanations from LLMs. Specifically, we denote the stereotypical bias explanations generated by LLMs and SLMs as the distribution P_l and P_s. Our objective function is $H(P_l, P_s) = \mathbb{E}_{y \sim P_l(y)} [-\log P_s(y)]$. Knowledge is transferred to SLMs by encouraging them to match the generations of LLMs.

4 Experiments

We design experiments to test the effectiveness of our methods toward answering three questions: **RQ1**: Does GPT-4o-mini make more accurate decisions than humans? **RQ2**: Can SLMs learn to generate explanations? **RQ3**: How GPT-4o-mini differ from human-generated decisions?

4.1 Measure Validation (RQ1)

Method We use **stereotype score** [4] to evaluate the performance of GPT-4o-mini and human annotations. Stereotype score is a continuous value from −1 to 1 used to indicate the stereotype of a sentence, with −1 indicating lower stereotypes and 1 indicating higher stereotypes. We chose RoBERTa version[2)] with the highest Pearson's r as our scoring model.

Table 2 Overall performance of training SLMs to generate explanations. **Bold** indicates the best performance.

Model	Faithful	BLEU	ROUGE	BERTScore
GPT-2 (124m)	77.76	9.48	20.55	85.36
OPT-125m	90.88	33.63	44.19	93.07
Bloomz-560m	81.77	13.19	25.87	88.66
OPT-350m	95.44	34.24	45.51	93.44
Phi-1.5 (1.3b)	72.38	12.54	25.15	88.66
OPT-1.3b	**98.20**	**36.20**	**47.45**	**93.93**

Results Figure 1 demonstrates the difference in stereotype scores between GPT-4o-mini and human annotations. Firstly, the human-annotated pro- and anti-stereotypical samples (blue and green) have closer stereotype scores than GPT-4o-mini (orange and red). In addition, GPT-4o-mini-annotated pro- and anti-stereotypical samples achieve the highest and lowest stereotype scores, respectively. This indicates that GPT-4o-mini are more correlated with stereotype scores than human annotations. Specifically, in the stereotype scores of the human-annotated samples, the median difference between pro- and anti-stereotypical samples is 0.015, whereas the corresponding difference for GPT-4o-mini-annotated samples is 0.079. This indicates that GPT-4o-mini exhibits a more significant ability to distinguish between pro- and anti-prototypical samples than humans. This also indicates that GPT-4o-mini may be more accurate than human annotation on highly subjective tasks. The results inspire future research on the usage of LLMs in stereotypical bias annotation. Please refer to Appendix A for specific bias types.

4.2 Explanation with SLMs (RQ2)

In this section, we train SLMs for stereotypical bias explanation.

Dataset We randomly split our dataset into 8:1:1 ratios for training, validation, and testing sets.

Models We use GPT-2 [9], Bloomz-560m [12], Phi-1.5 [13], and OPT [14] models as our baseline models. We download the weights and implementations of these models from the Huggingface library.[3)]

Metrics We train a binary classification model to evaluate the faithfulness [15] of explanations. Specifically, we collect 5,782 samples in the training set as positive samples and shuffle sentences and explanations to construct

2) https://huggingface.co/nlply/quantifying-stereotype-roberta

3) https://huggingface.co

5,782 negative samples. Then, we fine-tuned a RoBERTa model for the classification task, achieving 97.93% accuracy on the test set. In addition, we also use BLEU [16], ROUGE-L [17], and BERTScore [18] to evaluate the semantic completeness.

Results The experimental results are shown in Table 2. Firstly, OPT-1.3b gets the best performance on all metrics. Secondly, all SMLs achieve high faithfulness. However, except for OPT models, the semantic completeness of the explanations generated by the other models is relatively limited. In addition, OPT models get the best performance in the same parameter scale.

4.3 GPT-4o-mini vs. Human (RQ3)

Figure 2 shows the comparison of GPT-4o-mini with human annotations. We found differences between GPT-4o-mini and human-annotated samples, mainly in the counterfactual data (the orange bar). Specifically, the differences between GPT-4o-mini and human annotations are close to or greater than 50% on counterfactual data for all bias types. This indicates that human-rewritten counterfactual data, such as anti-stereotypical samples, may be unreliable. Moreover, GPT-4o-mini shows significant differences from human annotations in specific bias types, such as *nationality* and *physical-appearance*. This indicates that human-rewritten data may more likely introduce subjective judgments on specific bias types.

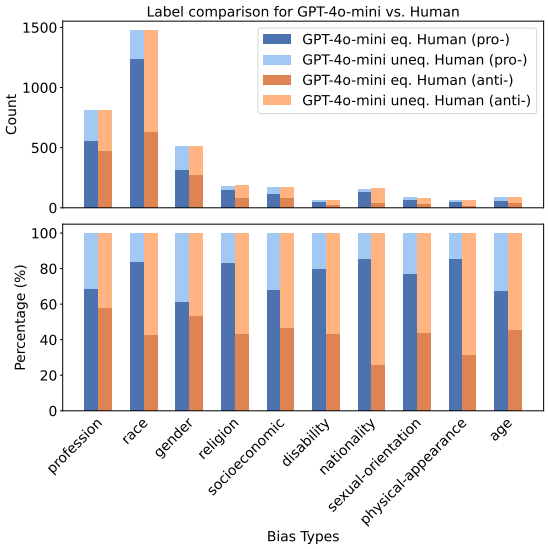

Figure 2 Comparison of GPT-4o-mini and human annotations on specific bias types. Blue indicates pro-stereotypical samples. Orange indicates anti-stereotypical samples. Deep indicates cases where GPT-4o-mini equal to human annotations. Light indicates cases where GPT-4o-mini unequal to human annotations.

5 Related Work

Stereotypical Biases in LLMs LLMs learn stereotypical human-like biases from human corpora. Nadeem *et al.* [1] and Nangia *et al.* [2] evaluated social biases in LLMs by constructing crowd-sourced datasets consisting of pro- and anti-stereotypical sentences. Subsequently, Blodgett *et al.* [3] indicated that these crowd-sourced datasets may not effective evaluated stereotypical biases in LLMs because of pitfall samples in these datasets. To mitigate the impact caused by pitfall samples on the evaluation, Liu [19] proposed to use the KL divergence of the Gaussian distributions as the evaluation scores. Furthermore, an overview and discussion of available datasets, evaluation methods, and debiasing methods is available in the survey by Gallegos *et al.* [20].

Explanation Generation Early explanation work [21] relied on supervised datasets to train explanation generators. Subsequently, Rajani *et al.* [22] proposed generating explanations or clarifications to improve task performance. Dalvi *et al.* [6] introduced entailment trees to explain answers to QA tasks. In recent studies, Marasovic *et al.* [23] study the effect of prompt format and model size on the plausibility of prompted explanations based on crowd-sourced worker annotations. Due to the excellent performance of GPT, Wiegreffe *et al.* [5] proposed to use GPT-3 to generate textual explanations for classification decisions. Their study revealed the great potential of LLMs in generating stereotypical bias explanations.

6 Conclusion

In this study, we use GPT-4o-mini to judge whether sentences exhibit pro- or anti-stereotypical biases and generate explanations. We find that GPT-4o-mini is more effective than human annotations in distinguishing pro- and anti-stereotypical bias, according to stereotype scores. In addition, SLMs can be trained to generate faithful explanations with our dataset. We also find that the main difference between GPT-4o-mini and human annotations is in the counterfactual data, and we point out that human-rewritten counterfactual data are unreliable. Our dataset will provide a valuable resource for studing generating stereotypical bias judgments and explanations with LLMs.

Acknowledgment

This work was supported by JST BOOST, Grant Number JPMJBS2407.

References

[1] 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 ACL-IJCNLP (Volume 1: Long Papers)**, pp. 5356–5371, Online, August 2021. Association for Computational Linguistics.

[2] 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 EMNLP**, pp. 1953–1967, Online, November 2020. Association for Computational Linguistics.

[3] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, **Proceedings of the 59th Annual Meeting of the ACL-IJCNLP (Volume 1: Long Papers)**, pp. 1004–1015, Online, August 2021. Association for Computational Linguistics.

[4] Yang Liu. Quantifying stereotypes in language. In **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1223–1240, 2024.

[5] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of NAACL: Human Language Technologies**, pp. 632–658, Seattle, United States, July 2022. Association for Computational Linguistics.

[6] Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on EMNLP**, pp. 7358–7370, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[7] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Proceedings of the 2022 Conference of NAACL: Human Language Technologies**, pp. 4602–4625, Seattle, United States, July 2022. Association for Computational Linguistics.

[8] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions. **arXiv preprint arXiv:2004.14546**, 2020.

[9] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[10] 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.

[11] Anthony G Greenwald and Linda Hamilton Krieger. Implicit bias: Scientific foundations. **California law review**, Vol. 94, No. 4, pp. 945–967, 2006.

[12] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. **arXiv preprint arXiv:2211.01786**, 2022.

[13] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: **phi-1.5** technical report. **arXiv preprint arXiv:2309.05463**, 2023.

[14] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[15] Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Jankowski, Yanghua Xiao, and Deqing Yang. Distilling script knowledge from large language models for constrained language planning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of ACL (Volume 1: Long Papers)**, pp. 4303–4325, Toronto, Canada, July 2023. Association for Computational Linguistics.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting on ACL**, ACL '02, p. 311–318, USA, 2002. Association for Computational Linguistics.

[17] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[18] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[19] Yang Liu. Robust evaluation measures for evaluating social biases in masked language models. In **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 38, pp. 18707–18715, 2024.

[20] 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, September 2024.

[21] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, **Advances in Neural Information Processing Systems**, Vol. 31. Curran Associates, Inc., 2018.

[22] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4932–4942, Florence, Italy, July 2019. Association for Computational Linguistics.

[23] Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. Few-shot self-rationalization with natural language prompts. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, **Findings of the Association for Computational Linguistics: NAACL 2022**, pp. 410–424, Seattle, United States, July 2022. Association for Computational Linguistics.

A Specific Bias Types

As shown in Figure 3, GPT-4o-mini outperforms human annotations on all bias types (higher stereotype scores for pro-stereotypical samples and lower stereotype scores for anti-stereotypical samples). Surprisingly, the human annotations exhibit negative correlations with stereotype scores in the *gender*, *socioeconomic*, *disability*, and *age* bias types. This indicates that humans face more significant challenges in rewriting samples of these bias types. In addition, GPT-4o-mini has a larger interquartile range (IQR) on anti-stereotypical samples in *physical-appearance* bias type. This indicates that GPT-4o-mini may have difficulty in judging *physical-appearance* bias type.

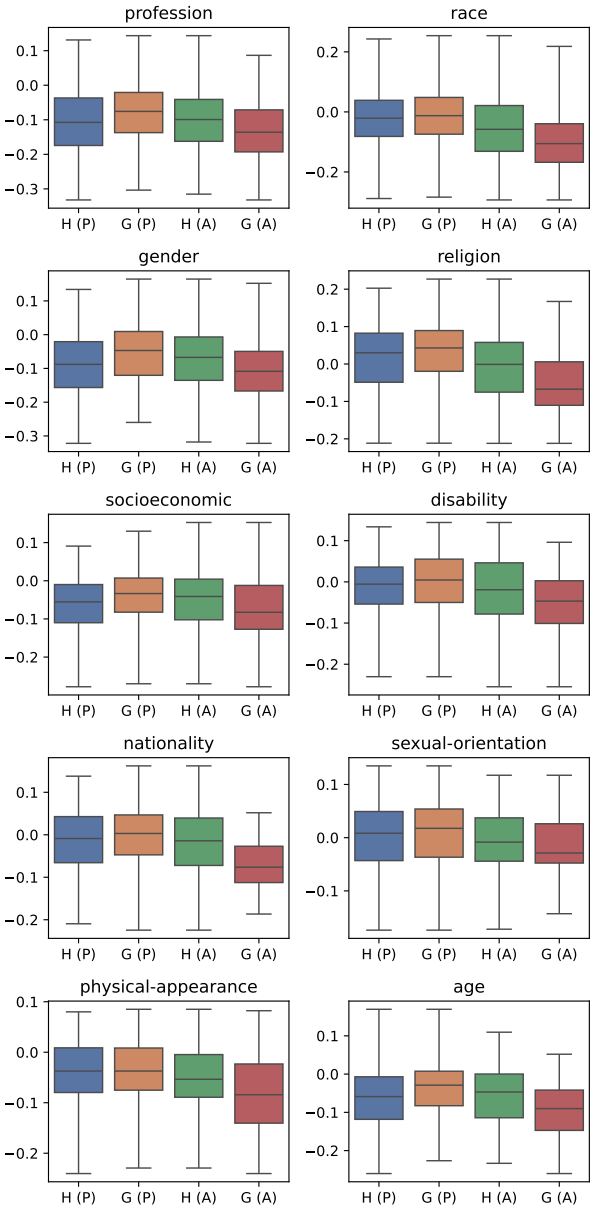

Figure 3 Boxplot of stereotype scores of GPT-4o-mini vs. human annotations on our dataset for specific bias types.