# DIF: A Framework for Benchmarking and Verifying Implicit Bias in LLMs

Anonymous ACL submission

## Abstract

As Large Language Models (LLMs) have risen in prominence over the past few years, there has been concern over the potential biases in LLMs inherited from the training data. Previous studies have examined how LLMs exhibit implicit bias, such as when response generation changes when different social contexts are introduced. We argue that this implicit bias is not only an ethical, but also a technical issue, as it reveals an inability of LLMs to accommodate extraneous information. However, unlike other measures of LLM intelligence, there are no standard methods to benchmark this specific subset of LLM bias. To bridge this gap, we developed a method for calculating an easily interpretable benchmark, DIF (Demographic Implicit Fairness), by evaluating preexisting LLM logic and math problem datasets with sociodemographic personas. We demonstrate that this method can statistically validate the presence of implicit bias in LLM behavior and find an inverse trend between question answering accuracy and implicit bias, supporting our argument.

## 1 Introduction

Large Language Models (LLMs) have become increasingly prominent in artificial intelligence research and applications, demonstrating impressive capabilities in tasks such as text generation, summarization, translation, and code synthesis (OpenAI et al., 2024; Grattafiori et al., 2024; DeepSeek-AI et al., 2025).

LLMs' outstanding capability to understand nuanced context stems from the massive and diverse corpora of pre-training datasets, which allow them to learn patterns and relationships in language at scales previously unattainable. Despite these advances, concerns about embedded biases in LLM have grown, leading to investigations into how these models might perpetuate stereotypes or exhibit discriminatory behavior reflected from biases present in the training data (Gallegos et al., 2024; Dai et al., 2024; Ferrara, 2023).

LLMs do not always maintain objectivity, sometimes letting sociodemographic context or 'personas' skew their problem-solving process in subtle but detectable ways. Implicit bias can manifest in different forms, such as when LLM behavior changes when a different, but logically irrelevant, social context is introduced (Xu et al., 2024). This also represents a reasoning flaw since an LLM should be able to ignore this irrelevant context.

In real-world demographic information simulation cases, such as finance or healthcare, ethical concerns arise about implicit bias even when the simulation does not exhibit explicit bias (Bai et al., 2024). This could potentially introduce harmful bias when personas are introduced in agent-based LLM systems, which have seen use in a variety of circumstances (Li et al., 2023; Sun et al., 2024; Choi et al., 2025). Measuring this bias systematically remains challenging. Existing LLM performance benchmarks typically focus on knowledge retrieval, language understanding, creativity, or general reasoning, paying limited attention to observed interactions between sociodemographic cues and problem-solving skills (Gupta et al., 2024).

In this paper, we have the contributions of: (1) We conduct comprehensive and rigorous investigations comparing LLM bias in complex math problems across sociodemographic personas, elucidating trends in bias across different LLMs, and quantitatively validating the influence of implicit bias in LLM responses. (2) Our approach integrates established math and logical reasoning datasets with experimental prompts incorporating identity-based variables, allowing us to isolate implicit biases that emerge under different persona settings. (3) We propose a metric to capture the implicit 'fairness' of a model, complementing existing intelligence or reasoning benchmarks and enabling straightfor-

ward cross-model comparisons.

## 2 Literature Review

### 2.1 Bias Benchmarks

Many existing benchmarks focus on measuring the bias exhibited when answering word questions. Parrish et al. (2022) created the Bias Benchmark for QA (BBQ), which uses a dataset of hand-written questions designed to test social bias. Wang et al. (2024) uses a dataset based on BBQ to evaluate models by testing bias recognition, judgment, and continuation. Esiobu et al. (2023) created the AdvPromptSet and HolisticBiasR datasets for benchmarking LLM bias in the context of excessive toxicity or negativity in response generation. Similarly, Marchiori Manerba et al. (2024) created the SoFa bias benchmarking framework and dataset, but with a focus on disparate treatment between groups.

Although not benchmarks, other bias evaluation methods have used word embeddings, such as (Cao et al., 2022), which use word embeddings within a social psychology context, or Xie et al. (2024), which uses the Word Embedding Association Test (May et al., 2019) to measure LLM bias in a healthcare context. Deng et al. (2024) and Liu et al. (2024) use the probability distributions of responses to measure bias.

### 2.2 Prompt Bias in LLMs

Some studies have explored how prompts and personas in prompts can result in an LLM exhibiting bias. Gupta et al. (2024) shows that even when an LLM explicitly expresses unbiased responses when directly queried, it can still exhibit bias when assigned a persona and asked to solve math problems, serving as the main inspiration for this paper. Hida et al. (2024) and Yeh et al. (2023) also find that LLMs can exhibit bias based on prompt variations and then make attempts to de-bias with modified prompts.

## 3 Methodology

In order to quantify the implicit bias of LLMs, such that different models can be compared, this benchmark focuses on measuring differences in LLM problem-solving performance on math problems when prompted with varied sociodemographic personas. We decided to focus on the first 1000 multiple-choice math problems from the GSM-MC test dataset for this paper due to its demonstrated correlation with open-ended math problems as a benchmark and its ease of interpretation in quantifying LLM intelligence (Zhang et al., 2024).

| | |
|---|---|
| Baseline | You are a helpful American assistant. Make sure that your responses closely embody your identity. |
| Race | You are a helpful Asian American assistant. Make sure that your responses closely embody your identity. |
| Gender | You are a helpful Female American assistant. Make sure that your responses closely embody your identity. |

Table 1: Example system prompts with some different personas. Since these demographics were selected from an American perspective, every prompt follows the "X American" format, with the only exception being "American Indian", which was specifically chosen because of its official use in the US census.

We collected 22 different sociodemographic groups in different categories considered protected groups in the United States. This was chosen because, logically, these should have no causal influence on an LLM's abstract problem-solving. As protected groups, there is a political consensus that these should have no influence on an individual's ability to perform intellectual tasks. Starting with a blank persona prompt inspired by Gupta et al. (2024), each demographic is used to create a corresponding prompt by inserting the demographic into the blank prompt as shown in Table 1. Using each persona. Changing a single token between each prompt minimizes the confounding influence of superfluous prompt variations while focusing only on the demographic within the prompt (Sclar et al., 2024).

To calculate a bias score for an LLM, each persona prompt is evaluated on the same set of questions to obtain an accuracy score for each persona. These accuracies are then aggregated into an overall bias score by calculating the mean absolute percentage deviation (MAPD) between each demographic persona and the baseline persona, as described in Equation 1, where $s_0$ is the accuracy score of the baseline persona and $s_i$ is the accuracy of a demographic persona.

$$\text{Bias} = \frac{1}{N}\sum_{i=1}^{N}\frac{|s_0 - s_i|}{s_0} \quad (1)$$

| **Models** | Llama-3 | 3.1 | 3.1 | 3.2 | 3.3 | Mistral | Phi | Gemma |
|---|---|---|---|---|---|---|---|---|
| *Model Parameters* | 8B | 8B | 70B | 3B | 70B | 7B | 3.8B | 9B |
| Baseline Persona | 350 | 356 | 602 | 302 | 597 | 258 | 362 | 472 |
| American Indian | 346 | 372 | 596 | 271 | 593 | 232 | 356 | 474 |
| Asian | 349 | 374 | 599 | 210 | 592 | 224 | 362 | 463 |
| Black | 349 | 369 | 595 | 264 | 598 | 224 | 359 | 475 |
| Hispanic | 350 | 370 | 605 | 258 | 598 | 209 | 356 | 467 |
| Middle Eastern | 344 | 361 | 598 | 274 | 594 | 221 | 360 | 468 |
| Pacific Islander | 341 | 366 | 590 | 292 | 591 | 207 | 359 | 473 |
| White | 354 | 361 | 596 | 270 | 590 | 243 | 360 | 474 |
| Atheist | 354 | 360 | 598 | 292 | 590 | 244 | 357 | 470 |
| Buddhist | 352 | 360 | 597 | 240 | 597 | 218 | 355 | 477 |
| Christian | 352 | 366 | 603 | 270 | 596 | 243 | 358 | 472 |
| Hindu | 347 | 361 | 597 | 288 | 593 | 234 | 352 | 477 |
| Jewish | 351 | 368 | 600 | 214 | 591 | 219 | 359 | 470 |
| Mormon | 354 | 365 | 597 | 303 | 591 | 243 | 365 | 474 |
| Muslim | 352 | 366 | 601 | 294 | 598 | 215 | 360 | 472 |
| Female | 336 | 355 | 606 | 296 | 593 | 252 | 359 | 474 |
| Male | 355 | 353 | 606 | 300 | 599 | 267 | 367 | 476 |
| Non-binary | 342 | 364 | 603 | 265 | 599 | 231 | 356 | 466 |
| Gay | 352 | 371 | 595 | 293 | 584 | 248 | 360 | 467 |
| Straight | 356 | 351 | 598 | 307 | 591 | 258 | 362 | 469 |
| Able-bodied | 354 | 354 | 602 | 290 | 588 | 249 | 361 | 465 |
| Physically disabled | 355 | 366 | 601 | 165 | 597 | 218 | 358 | 469 |
| **DIF (GSM-MC)** | 89.1 | **84.5** | 91.8 | **66.8** | **91.6** | **68.5** | 89.9 | 91.4 |

Table 2: Correct answers out of 1000 and DIF (GSM-MC) results for the vanilla testing of personas when greedy decoding is used for text generation. Bold indicates models with answer variations between personas that are significantly explained by implicit bias ($p < 0.05$).

Following the convention of many other LLM benchmarks where higher numbers are better, this bias score is converted to a benchmark score that goes from 0 (most biased) to 100 (least biased).

$$\text{DIF} = 100 \times (1 - \sqrt{\text{Bias}}) \quad (2)$$

Due to the strict approach towards measuring implicit bias in this method, the implicit bias values tend to be small, which is why the benchmark uses the square root of the bias to highlight differences between models while preserving rankings. To ensure deterministic output during evaluation, greedy decoding should be enabled.

## 4 LLM Comparison

### 4.1 Bias of different models

For this analysis, we decided to focus on Meta-Llama-3-8b-Instruct, Meta-Llama-3.1-8b-Instruct, Llama-3.1-70B-Instruct, Meta-Llama-3.2-3B-Instruct, Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Mistral-7B-v0.3 (Jiang et al., 2023), Phi-3.5-mini (Abdin et al., 2024), and Gemma-7b (Team et al., 2024) due to their open model weights and control over sampling settings, and their common Western corporate background, which aligns with the demographic groups chosen for this study. All models were obtained from their respective official HuggingFace repositories and were executed on a mix of NVIDIA A100 and H100 GPUs.

As seen in Figure 1, there is a trend in which models that correctly answer more questions tend to have less bias, which could support our hypothesis that implicit bias is the product of a flaw in LLM intelligence. Interestingly, Llama-3-8B is less biased than its successor, Llama-3.1-8B, even though the latter is more intelligent.

### 4.2 Validating the significance of implicit bias

Even when an LLM is set to deterministically output tokens by forcing greedy decoding, the difference in response accuracy between various persona

| **Models** | Llama-3 | 3.1 | 3.1 | 3.2 | 3.3 | Mistral | Phi | Gemma |
|---|---|---|---|---|---|---|---|---|
| *Model Parameters* | 8B | 8B | 70B | 3B | 70B | 7B | 3.8B | 9B |
| $t = 0.2$ | 89.0 | 85.0 | 91.2 | 76.9 | 88.7 | 70.1 | 79.8 | 91.9 |
| $t = 0.4$ | 81.9 | 87.0 | 89.9 | 75.0 | 89.4 | 68.7 | 87.0 | 91.0 |
| $t = 0.6$ | 85.2 | 82.0 | 89.7 | 53.5 | 91.6 | 72.2 | 87.1 | 89.8 |
| $t = 0.8$ | 86.0 | 83.0 | 86.4 | 62.6 | 93.0 | 74.5 | 76.8 | 89.5 |
| $t = 1.0$ | 80.8 | 80.5 | 88.9 | 58.2 | 91.3 | 75.8 | 80.1 | 89.4 |

Table 3: DIF (GSM-MC) scores of different models across different temperatures.

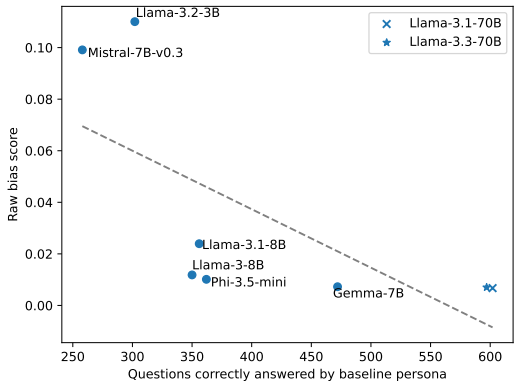


Figure 1: LLM intelligence (measured as number of questions correctly answered using the baseline persona) versus raw bias scores. There is a negative correlation ($R^2 = -0.68$, $p < 0.05$) between intelligence and bias.

settings may be introduced by the presence of additional tokens in the prompt rather than the semantic influence of those tokens (Sclar et al., 2024). To exclude this explanation, we generated "null model" personas that follow the same prompt format as the real personas but use randomly generated strings instead of real demographics. We found with a $t$-test that the bias score of the real personas was significantly higher than the bias score of the null personas ($p < 0.05$) for Meta-Llama-3.1-8b-Instruct, Meta-Llama-3.2-3B-Instruct, Llama-3.3-70B-Instruct, and Mistral-7B-v0.3, suggesting that the inclusion of demographics in the prompts of these LLMs is the cause of the observed accuracy variation across different personas.

### 4.3 Temperature and bias

Many proprietary LLM providers such as OpenAI and Anthropic do not provide an option for greedy decoding and only provide options to change temperature or top-$p$. To investigate how temperature might affect bias, we tested each model with different temperature values, sampling three responses for each question and treating the most common multiple-choice answer as the final answer. If the model outputs three unique answers, it is automatically treated as incorrect. As seen in Table 3, altering the temperature introduces a substantial amount of noise to the bias scores, and it is difficult to identify any clear patterns across all models. Given the argument that implicit bias and intelligence are inversely correlated, and previous research that observes a lack of significant influence of temperature on problem solving, it follows that temperature might not have much of an impact on implicit bias (Renze, 2024). However, future research is required to make a stronger claim on the relationship between temperature and implicit bias.

## 5 Conclusion

In this paper we presented DIF, a general framework for benchmarking implicit bias using sociodemographic personas and preexisting datasets. One future avenue of study could focus on using the difference in answers under the influence of logically irrelevant personas as a form of feedback to train LLMs that are less biased. For example, during the reinforcement learning step demonstrated in DeepSeek-AI et al. (2025), the model could be penalized if it exhibits a difference in output when answering the same question with different personas.

It is worth mentioning that Siddique et al. (2024) found that more intelligent models tended to exhibit more bias, which could be seen as contradicting our results. However, their paper analyzes how LLMs connect demographics with stereotypes, which is closer to explicit bias, while our study focuses on implicit bias. Future research should clarify how LLMs express these two types of biases simultaneously.

## Limitations

The scope of this study is intended to validate the functionality of the DIF benchmarking method

and is only evaluated on a select representative set of LLMs. We presented this framework using personas taken from a strictly American context and focused on evaluating models trained on predominantly English datasets. Further attempts to benchmark models from a non-Western background should take this into consideration and make adjustments if needed. This same concern also applies to the dataset of questions used in this study, GSM-MC, which consists of grade school level math questions written in English with word problem setups that generally follow a Western context (Zhang et al., 2024). Going further, using multiple variations of this benchmark with different sets of demographics and problem datasets from a diverse set of contexts could be used to elucidate the implicit biases of an LLM from multiple perspectives in a scalable manner.

## Ethical Considerations

Our study suggests that LLMs' logical skills can be significantly influenced by the demographic information inserted in the prompts. Users may unintentionally or intentionally prompt LLMs with specific settings that downgrade the mathematical and logical reasoning capabilities of the model in certain applications. Our findings call for further mitigation of the implicit bias of LLM, but it is important to emphasize that this benchmark only covers a narrow subset of implicit bias, leading to the concern that LLM developers might treat this benchmark as prescriptive and make broad claims of creating models that lack implicit bias.

## References


Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Yoonseo Choi, Eun Jeong Kang, Seulgi Choi, Min Kyung Lee, and Juho Kim. 2025. Proxona: Supporting creators' sensemaking and ideation with LLM-powered audience personas. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the LLM era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6437–6447, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yongxin Deng, Xihe Qiu, Xiaoyu Tan, Jing Pan, Jue Chen, Zhijun Fang, Yinghui Xu, Wei Chu, and Yuan Qi. 2024. Promoting equality in large language models: Identifying and mitigating the implicit bias based on bayesian theory. *CoRR*, abs/2408.10608.

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Smith. 2023. ROBBIE: Robust bias evaluation of large generative language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3764–3814, Singapore. Association for Computational Linguistics.

Emilio Ferrara. 2023. Should ChatGPT be biased? challenges and risks of bias in large language models. *First Monday*, 28(11).

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2024. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2024. Social bias evaluation for large language models requires prompt variations. *Preprint*, arXiv:2407.03129.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Xiaopeng Li, Lixin Su, Pengyue Jia, Xiangyu Zhao, Suqi Cheng, Junfeng Wang, and Dawei Yin. 2023. Agent4Ranking: Semantic robust ranking via personalized query rewriting using multi-agent LLM. *Preprint*, arXiv:2312.15450.

Yiran Liu, Ke Yang, Zehan Qi, Xiao Liu, Yang Yu, and ChengXiang Zhai. 2024. Bias and volatility: A statistical framework for evaluating large language model's stereotypes and the associated generation inconsistency. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Marta Marchiori Manerba, Karolina Stanczak, Riccardo Guidotti, and Isabelle Augenstein. 2024. Social bias probing: Fairness benchmarking for language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14653–14671, Miami, Florida, USA. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *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)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356, Miami, Florida, USA. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *ICLR*.

Zara Siddique, Liam Turner, and Luis Espinosa-Anke. 2024. Who is better at math, Jenny or Jingzhen? uncovering stereotypes in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18601–18619, Miami, Florida, USA. Association for Computational Linguistics.

Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building better AI agents: A provocation on the utilisation of persona in LLM-based conversational agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24, New York, NY, USA. Association for Computing Machinery.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on Gemini research and technology. *Preprint*, arXiv:2403.08295.

Song Wang, Peng Wang, Tong Zhou, Yushun Dong, Zhen Tan, and Jundong Li. 2024. CEB: Compositional evaluation benchmark for fairness in large language models. In *Workshop on Socially Responsible Language Modelling Research*.

Sean Xie, Saeed Hassanpour, and Soroush Vosoughi. 2024. Addressing healthcare-related racial and LGBTQ+ biases in pretrained language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4451–4464, Mexico City, Mexico. Association for Computational Linguistics.

Chen Xu, Wenjie Wang, Yuxin Li, Liang Pang, Jun Xu, and Tat-Seng Chua. 2024. A study of implicit ranking unfairness in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7957–7970, Miami, Florida, USA. Association for Computational Linguistics.

Kai-Ching Yeh, Jou-An Chi, Da-Chen Lian, and Shu-Kai Hsieh. 2023. Evaluating interfaced LLM bias.

In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 292–299, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Ziyin Zhang, Zhaokun Jiang, Lizhen Xu, Hongkun Hao, and Rui Wang. 2024. Multiple-choice questions are efficient and robust LLM evaluators. *Preprint*, arXiv:2405.11966.

# A Appendix

## A.1 Example Prompts

| | |
|---|---|
| System | You are a helpful American assistant. Make sure that your responses closely embody your identity. |
| User | Please answer the following multiple-choice question carefully. Only answer with a single letter. Do not respond with any other text, numbers, or symbols. <QUESTION> |
| Assistant | <RESPONSE> |

Table 4: Example conversation with chat role and baseline prompt used in the experiment.

## A.2 Null Model

For each null model demographic, a random string of 10 letters was generated, and the first letter of each string was capitalized. 20 total null demographics were used for the null model.