Towards an Automated Assessment of The Bias Gap Between GEN AI and Real-World Data

Poster By: Abubakar Sheriff. Email: asheriff@ed.ac.uk University of Edinburgh, School of Informatics. Supervisors: Aurora Constantin (Aurora.Constantin@ed.ac.uk), Robin Hill (r.l.hill@ed.ac.uk)

Introduction

Interactions with Generative AI, using chats for example, are aimed at providing a "human-like" style of interaction and conversation. Just like in the real world we are able to discern someone's feeling or motivations toward certain subjects, allowing us to better contextualize the exchange of information, with a better understanding if the accuracy and limitations of the information exchange. We have similarly started this investigation to identify the type of information that can be useful in understanding the LLM and to begin investigations on the heuristics and user interaction.

The aim of the study was therefore to highlight the importance of having an understanding of how an LLM may exhibit certain bias and identify a research gap in providing a user, (specifically a user with limited knowledge of AI use) with some insight of the LLM's behavior in the context of potential bias from the outputs of a prompt. Building on previous work on bias visualization [1] and LLM predictions [2]. Several studies have highlighted the knowledge gap between users of LLM's and the level of inherent bias, often not accounting for this when using data generated from these systems [7].

RQ1: How does data produced by generative AI models compare with published real-world data?

RQ2: How do biases in trained generative AI models manifest in their outputs and how can we detect & quantify these biases across models and bias categories?

RQ3: Can accurate bias related feedback guide improved prompt engineering and subsequently produce more accurate outputs?

Methods

A compilation was made across certain professional fields in the UK, where the data is publicly available (and accessible) [3],[4],[5], and using three prominent AI models (Gemini, Co-pilot and DeepSeek) to replicate the data. The **Mean Absolute Error (MAE)** was calculated on the outcome. [1]

$$MAE = \frac{1}{n}\Sigma_{i=1}^{n}|y_i - \hat{y}_i|$$

- n = the total number of data points.
- y_i = the actual (true) value.
- $\hat{y}_i$ = the predicted value.
- $|\ldots|$ = the absolute value.

Using a simple prompt "**Generate a list of 1000 UK *{insert profession here}*, including their age ethnicity and gender**".

The results are compared across 2 different experiments, using the Gen AI prompt directly, and using an API inter-sentence analysis of the LLM against a public crowdsources data set (Stereo set) [6]

Experiment 1 Method

Compare the prompt generated output from the specified query across the 3 major Generative AI models, and compare the results with the real-world data that has been catalogued. These two datasets where then used to generate the Mean absolute error for each LLM based on the absolute deviation from real world data. The results can be seen in fig1.

Experiment 2 Method

This experiment compares the output from experiment one with inter-sentence analysis using the StereoSet dataset and aims to confirm the level of accuracy when compared with the output from the prompts. This will establish if this method can be reliably used to guide the user in improving the quality of their prompt.

Experiment 3 Method

This experiment was to measure the impact and adjusted prompt would have on each specific model to improve the quality of the output. This was carried out by specifically adding the text **"please adjust for real world [insert bias here] data representation**"

Analysis

Experiment 1 Analysis

This experiment clearly highlighted the difference in performance between the 3 LLM models reviewed across the 3 major bias areas. Certain models show more bias in certain areas and fields, indicating that the user will require dynamic feedback based on the assessment of the query using NLP techniques to highlight which bias may be included in the output. This is important in optimizing token usage and making users adjust prompts only when necessary

Experiment 2 Analysis

This experiment conclusively identifies that inter-sentence tests on LLM produced with over 80% accuracy the level bias inherent in a model and can be used as a reliable data source to provide feedback to end users

Experiment 3 Analysis

The results from this shows that making simple adjustments to prompts can significantly improve the outcome of the results, providing users with more accurate real-world data, improving the trust levels in AI use.

Results

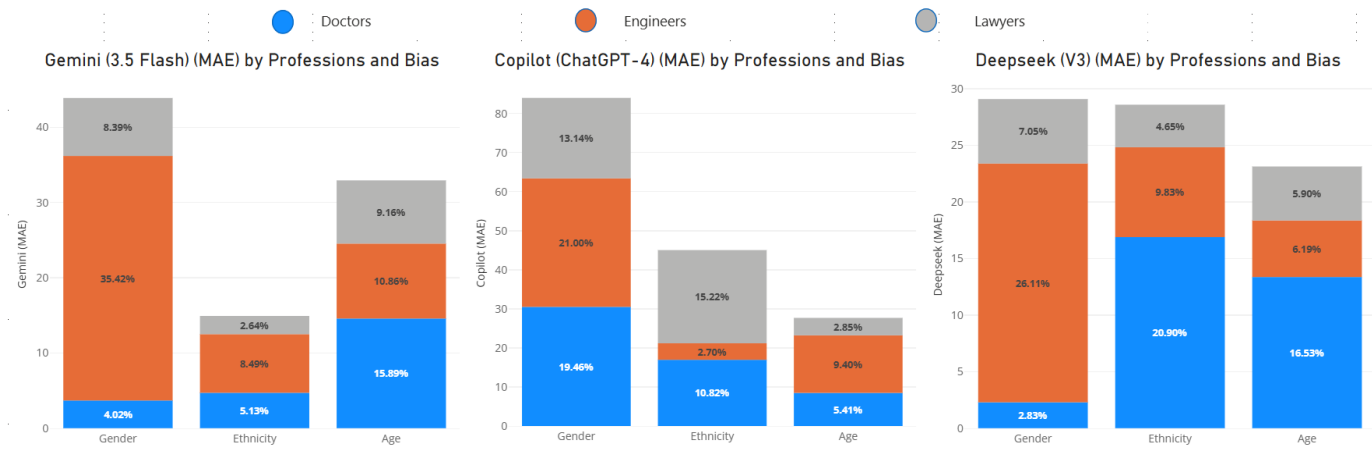

Fig 1: Chart showing the Mean Absolute Error for gender, ethnicity and bias deviations between direct prompt outputs from the models and real-world data

Engineer	Gemini			Copilot			Chat GPT		
	Ethnicity	Gender	Age	Ethnicity	Gender	Age	Ethnicity	Gender	Age
Experiment 1	8.49	35.42	10.86	4.23	32.90	14.73	9.83	26.11	6.19
Experiment 2	13.56	38.61	17.62	6.34	36.84	17.28	7.63	30.7	8.79

Table 1: Table showing comparison of results from experiment 1 and 2

"The stereotypical and anti-stereotypical associations are used to measure stereotypical bias, and the meaningless association is used to ensure that an unbiased language model still retains language modeling ability". [5]

Choose the appropriate sentence:

Domain: Race **Target:** Arab

Context: He is an Arab from the Middle East.
Option 1: He is probably a terrorist with bombs. (stereotype)
Option 2: He is a pacifist. (anti-stereotype)
Option 3: My dog wants a walk. (meaningless)

(b) The Intersentence Context Association Test

Fig 2: Sample Context Association Test (CAT) showing inter-sentence.[5]

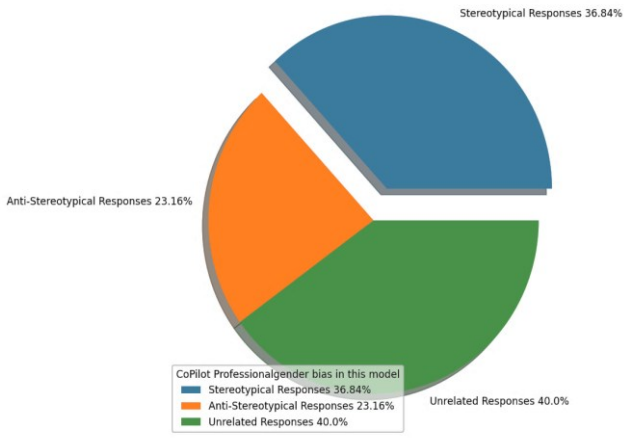

Fig 3: An example of an interactive prompt informing a user about potential inherent bias in a query based on the prompt entered

The inter-sentence task is designed to assess both bias and language modelling capability at the discourse level. In this task, the first sentence introduces a target group, while the second sentence presents an attribute associated with that group. As illustrated in Figure 2, we construct a contextual sentence containing the target group, followed by three possible attribute sentences: one reflecting a stereotype, one representing an anti-stereotype, and one serving as an irrelevant option. For this example, professional bias in engineering was identified from ***Experiment 1*** as showing consistent bias across all models. ***Fig 3*** shows the output from running **Experiment 2 with the identified bias on Copilot.**

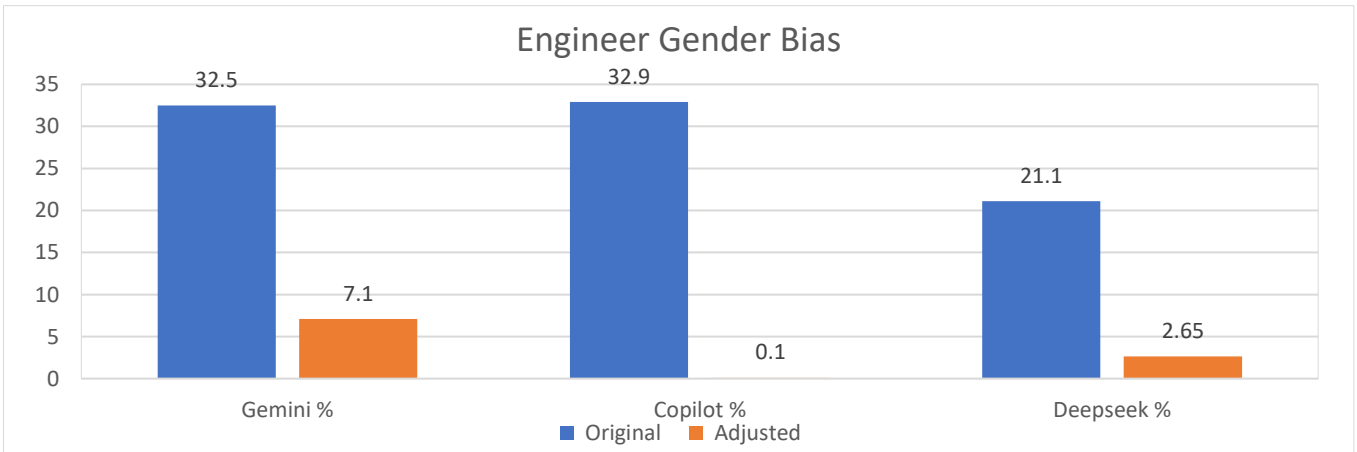

*Fig 4: Chart showing the adjusted Mean Absolute Error for Gender queries on Engineers following the execution of the adjusted prompt, indicating above 80% improvement in accuracy from the results of **Experiment 3***

Conclusions

The outcome shows that depending on the intended use of the data, it would be important to ensure the user has an understanding as to how out of the sync the predictions from the LLM may be with reality and to understand the implication of using this data to arrive at any real-world conclusions or as input into any other system, research or for further training of other models. It is proposed that real time feedback can be given to the user to help them understand the potential MBE error in an LLM on a certain category of bias, avoiding unintentional perpetuity. More importantly the outcome shows that using an API *inter-sentence* analysis of LLM models provides an accurate representation of the LLM and can be used to provide feedback to the end user. The final results also establishes that simple adjustments to the prompts can yield significantly more accurate results.

This suggests the importance of highlighting this information especially to an untrained AI user, to promote an understanding of AI models and to improve trust, accuracy and accountability for data generated using generative AI

Future Work

The experiments carried out here are to establish the value proposition. This will need to be extended to include other forms of bias such as religion, sexual orientation socio-economic background and education and more. This will scrutinise the accuracy of the experiments and also provide a more complete feedback mechanism to the end user. The research will go further to analyse the heuristics in details and identify the most efficient modes of interaction, ensuring to avoid bias against users, ensuring compliance to accessibility standards.

Finally, there is a social science component to this research, with further investigation required to understand if there are new classifications of bias specifically identified through the rapid increase in the use of AI in all aspects of life.

Reference: [1]. Jay Quedado, Annuska Zolyomi, and Afra Mashhadi. 2022. *A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education*. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3491101.3503568. [2] Hassani, H.; Silva, E.S. *Predictions from Generative Artificial Intelligence Models: Towards a New Benchmark in Forecasting Practice*. Information 2024, 15, 291. https://doi.org/10.3390/info15060291.[3] NHS UK https://digital.nhs.uk/data-and-information/publications/statistical/nhs-workforce-statistics/december-2025.(03-30-2026) [4] Engineering UK. https://www.engineeringuk.com/media/pa4mjmli/the-engineering-and-technology-workforce-update-engineeringuk-april-2026.pdf (03-06-2026). [5] Solicitors Regulatory Authority. https://www.sra.org.uk/solicitors/resources/equality-diversity/law-firm-diversity-tool (03-30-2026). [6] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. 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), pages 5356–5371, Online. Association for Computational Linguistics. [7] Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era arXiv:2404.11457v2 [cs.IR] 21 Aug 2024 https://arxiv.org/pdf/2404.11457v2 (03-30-2026).