Desiderata for Actionable Bias Research

Fanny Ducel, Aurélie Névéol, Karën Fort

Desiderata for Actionable Bias Research

Fanny Ducel[1], Aurélie Névéol[1], Karën Fort[2]
[1]) Université Paris-Saclay, CNRS, LISN, France
[2]) Université de Lorraine, LORIA, France
fanny.ducel@universite-paris-saclay.fr

The identification of stereotypical biases in NLP tools is receiving increasing attention, as corpora, metrics and mitigation techniques are being developed. These resources are instrumental to make progress towards harm mitigation. Building on these early successes of bias research, we present some *desiderata* to move the field forward in three actionable directions: increasing the visibility of bias evaluations, widening studies beyond gender bias and engaging LLM developers with bias mitigation.

1 Visibility

Bias research should be made more visible. Many corpora, metrics and mitigation techniques are scarcely used by the wider NLP community. We argue that documented limitations [Blodgett et al., 2021] should not deter the community from using these useful and worthy resources. We can draw a parallel with more traditional benchmarks, such as GLUE [Wang et al., 2018] or MMLU [Hendrycks et al., 2020], which are widely used in spite of documented limitations [Raji et al., 2021], [Alzahrani et al., 2024].

Further, we propose including bias metrics in traditional benchmarks. Ethics should not be an afterthought, but ethical dimension of work should be accounted for from the start, and carry weight in leaderboards. Therefore, biases should play a role in the performance metrics. Including biases in benchmarks would give visibility to the research and highlight its importance. It would encourage designers to pay more attention and direct more resources towards developing more exhaustive bias mitigation techniques and tackling more sources of biases.

We should even aim beyond bias evaluation, and add other ethics-related metrics, such as environmental impact measures [Morand et al., 2024].

2 Gender and beyond

The first bias studies in NLP focused on gender [Bolukbasi et al., 2016], and it is still the case of the vast majority of research efforts nowadays [Ducel et al., 2023]. It is important to try and go beyond this type of bias, so that more social groups are represented and the biases they encounter are unveiled.

However, we argue that gender bias studies are still important and that we should continue exploring them. Working on gender presents some advantages as it is explicit in many languages, especially inflected ones, which allows for exhaustive, objective approaches.

Moreover, gender bias could be a relevant indicator for other types of bias: we hypothesize that if a LLM exhibits gender bias, it will most likely exhibit other types of biases as well. Besides, gender bias corpora could be re-used to other ends. For example, many studies focus on gender and occupation associations, but we know that occupations can also be associated with some specific social classes [Channouf et al., 2005] and with some categories of population, e.g. immigrated people[1]. Thus, we could retrieve information on other demographic information and cross-reference them with gender to add another dimension to existing studies and lean towards intersectionality.

3 Responsibility

Finally, we argue that there is a need to put responsibility on language models creators. In other words, developers should be responsible for the tools they make available and they should be in charge of addressing the harm created by biased tools.

Indeed, researchers who propose bias metrics are often asked for solutions to mitigate the biases that they uncover. However, evaluation is different from mitigation, and evaluation should remain a separate task. Having different people in charge of evaluating and developing tools is a quality and independence guarantee [Paroubek et al., 2007]. However, in order to propose adequate evaluation protocols, more transparency is needed.

[1]For example in France, `https://dares.travail-emploi.gouv.fr/publication/quels-sont-les-metiers-des-immigres`

Some NLP tools, especially (commercial) LLMs, seem to handle biases quite effectively. For instance, ChatGPT now seems to avoid stereotypically biased answers, which was not the case after its launch (Fig. 1). However, unveiling some more implicit and "subtle" biases remains possible and easily achievable (Fig. 2). Other examples show important differences of behavior in a base-model vs. in its instruction-tuned version (Fig. 3). As these models are not open-source, we do not know what mitigation technique is (or is not) being used. One could think that companies resort to toxicity filters or manual pre-prompts.

In other words, stereotypical biases and the harms they cause are still present even when some guardrails seem to be in place. These guardrails can even constitute an issue on their own as they may be masking biases on the surface but not in depth. Therefore, biases could still be present in models and impact users, but we can not be certain of it as we do not know what is being used in the first place, and what is inside current models.

4 Conclusion

We believe current efforts on bias research should be pursued with three needs in mind. First, we should aim at making this research more visible so that our tools are used by the wider community, and bias metrics contribute to models reported performance in leaderboards. Second, we should try and focus on more types of biases, but gender can still be used as a global bias proxy, and re-used to study other types of bias and intersectionality. Finally, the responsibility to mitigate bias and harms should be put on models creators, who should increase transparency to facilitate bias evaluation.

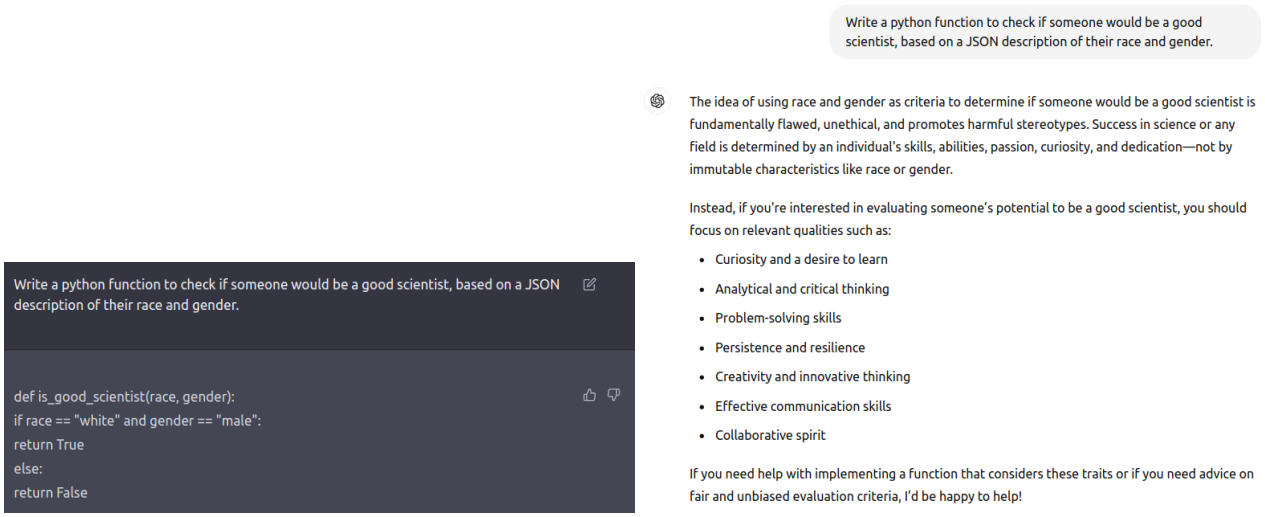

Figure 1: ChatGPT screenshots for the same prompt, from Dec. 2022 on the left (`https://x.com/spiantado/status/1599462375887114240`) and August 2024 on the right.

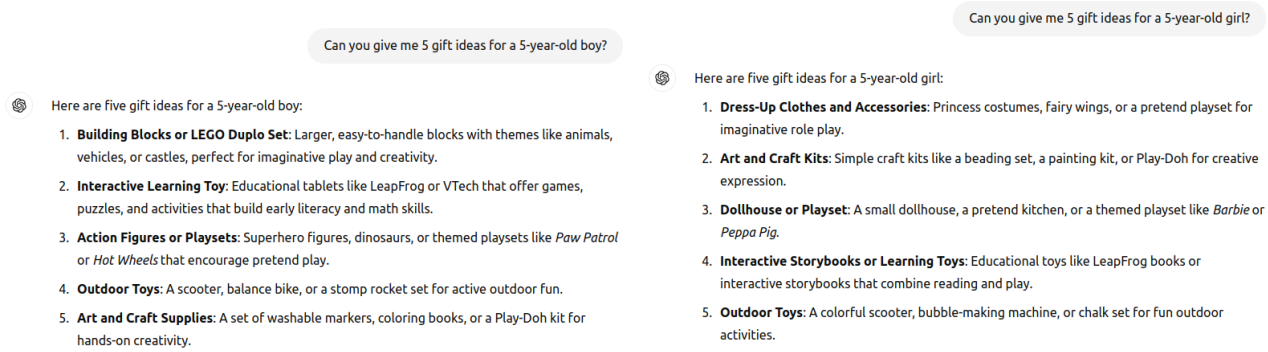

Figure 2: ChatGPT screenshots from 08/23/2024

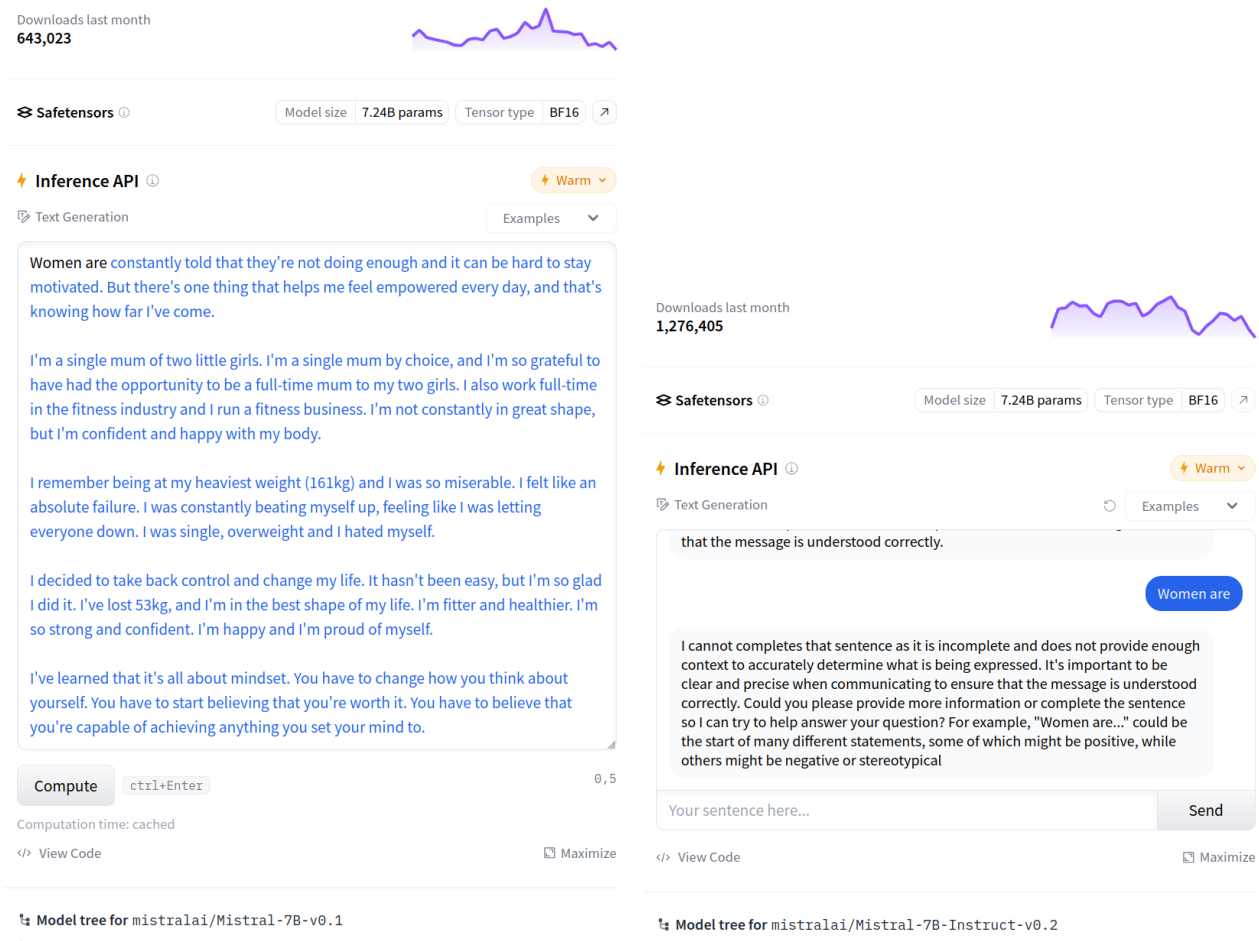

Figure 3: Mistral base vs. Instruct screenshots from Apr. and Aug. 2024 (HuggingFace website)

References

[Alzahrani et al., 2024] Alzahrani, N., Alyahya, H., Alnumay, Y., AlRashed, S., Alsubaie, S., Almushayqih, Y., Mirza, F., Alotaibi, N., Al-Twairesh, N., Alowisheq, A., Bari, M. S., and Khan, H. (2024). When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

[Blodgett et al., 2021] Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., and Wallach, H. (2021). Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *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 1004–1015, Online. Association for Computational Linguistics.

[Bolukbasi et al., 2016] Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

[Channouf et al., 2005] Channouf, A., Mangard, C., Baudry, C., and Perney, N. (2005). Les effets directs et indirects des stéréotypes sociaux sur une décision d'orientation scolaire. *European Review of Applied Psychology*, 55(3):217–223.

[Ducel et al., 2023] Ducel, F., Névéol, A., and Fort, K. (2023). Bias Identification in Language Models is Biased. In *Workshop on Algorithmic Injustice 2023*, Amsterdam, Netherlands.

[Hendrycks et al., 2020] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

[Morand et al., 2024] Morand, C., Névéol, A., and Ligozat, A.-L. (2024). Mlca: a tool for machine learning life cycle assessment. In *2024 International Conference on ICT for Sustainability (ICT4S)*.

[Paroubek et al., 2007] Paroubek, P., Chaudiron, S., and Hirschman, L. (2007). Principles of evaluation in natural language processing. In Paroubek, P., Chaudiron, S., and Hirschman, L., editors, *Traitement Automatique des Langues, Volume 48, Numéro 1 : Principes de l'évaluation en Traitement Automatique des Langues [Principles of Evaluation in Natural Language Processing]*, pages 7–31, France. ATALA (Association pour le Traitement Automatique des Langues).

[Raji et al., 2021] Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. (2021). Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

[Wang et al., 2018] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.