Low-Resource Fairness for Large Language Models: Black-Box Bias Evaluation and Mitigation

Luca Benedetto

January 2026

1 Team

- Supervisory team:
  - **Luca Benedetto** (luca.benedetto@telecom-sudparis.eu),
  - Amel Bouzeghoub
- Affiliated School of thesis supervisor: Telecom SudParis
- Thesis hosting team: ACMES, Samovar lab

2 Keywords

- AI Fairness and Safety
- Natural Language Processing
- Low-resource Fairness
- Bias Mitigation
- Accessible AI Safety

3 Introduction

Recent years witnessed a proliferation of AI models: modern Large Language Models (LLMs) proved very effective to target different tasks, and there is now a tendency to use them in a variety of domains, including education [11, 2], healthcare [12], and recommender systems [5, 20]. Crucially, some commercial models are available for free or for a (small) subscription fee and do not require technical knowledge. Their widespread availability, often through low-friction

interfaces requiring no technical expertise, has accelerated adoption in uncontrolled and non-standardized settings (e.g., students practising with ChatGPT). Even though this low barrier has potential benefits, ample research documented the biases exhibited by AI models ([7, 8, 13] among others), which can have implications across application domains (job recommendations [15], resume screening [19], education [18], inter alia).
While this lowered barrier of entry is central to the democratisation of AI, numerous studies have documented systemic and emergent biases in LLMs. To address these issues, the field of AI fairness has evolved in recent years, and there are now methods for bias detection and mitigation which are somewhat effective [9]. Still, the definition of *fairness* can vary across domains, and models that can be considered safe on one task might behave differently on others. Moreover, most models align with English-centric norms and risks, which poses questions about the cultural alignment of mitigation strategies [14, 1]. Bias detection and mitigation should thus be performed on each task before using pre-trained models, but this comes with an associated cost: fairness is *not* computationally free [4], and current state of the art bias detection and (especially) mitigation methods are resource intensive [16, 6, 9]. Unfortunately, the vast majority of AI adopters have limited resources, thus being unable to implement comprehensive bias evaluation/mitigation pipelines and relying on the mitigation performed by the companies and research labs training commercial and open weights-models. In practice, this contradicts the narrative of AI democratisation: while access to AI models is indeed being democratised, access to AI safety is not – thus creating a *fairness divide*.

4 Scientific goals

This project directly targets the fairness divide, with *the objective of developing low-resource bias evaluation and mitigation techniques, which can be used by small players without massive budgets to ensure the democratisation of AI Fairness*. This research defines low resource not by the availability of data for a specific language, but by a set of technical constraints: minimal computational cost, minimal model access, minimal (if any) model modifications, and minimal human labour. The innovative nature of this research lies in the focus on low-resource environments, and this three-way trade-off between mitigation cost, bias reduction, and task accuracy – while most of previous research focused on only two of these aspects. This project will:

- Study the feasibility of low-cost black box evaluation methods, and compare their faithfulness with more computationally expensive alternatives.
- Quantify the performance-per-compute-cost curve for bias mitigation across different tasks.
- Identify the trade-off between mitigation cost, bias reduction, and task accuracy, for low-cost mitigation techniques.

5 proposed approach / expected results

Primary outcome will be a low-resource fairness toolkit, providing the tools to perform low-resource bias evaluation and mitigation on a given domain and tasks. This tool will enable AI adopters to understand whether a given model can be used safely and effectively in their domain, requiring limited data and computation.

We will primarily focus on black-box models, using both commercial models (e.g., those from OpenAI, Anthropic, Google), which represent the most common AI adoption scenario, and open-weight models (without looking at their internal weights). In this setting, there is no access to the models beyond API-based queries, and the internal states of the model (e.g., weights and activations) are inaccessible, which is a significant obstacle for AI audits since it prevents deeper analysis [3]. This research aims to circumvent this issue taking inspiration from psychometrics: we hypothesise that the biases exhibited by the models can be modelled as *latent traits*. In other words, similar to how testing theories (such as Item Response Theory [10]) estimate the skill level of a learner by observing the correctness of their responses to exam questions, we will quantify the latent trait representing the bias exhibited by a model by observing its responses to different requests. Low-resource evaluation can then be performed in an analogous way to Computerised Adaptive Testing [17]: by picking the "right" questions, it is possible to obtain an accurate measurement of the latent trait with a fraction of the number of responses.

Then, we will be study low-resource mitigation techniques, focusing on training-free methods, that can be implemented during inference or as external modules, thus offering a lightweight and accessible approach to bias mitigation and AI safety (in contrast to computationally expensive alignment techniques which alter the model's weights). We will experiment on a variety of techniques at both pre-inference (the input to the model is automatically edited to minimise the risk of biased responses) and post-inference (the text generated by the LLM is automatically edited before delivering it to the end user) stage.

6 Future Prospects

Once completed, this low-resource fairness toolkit has the potential to support AI adopters in ensuring that they use these models safely, thus working towards the *democratisation of safe AI*, instead of the democratisation of unsafe AI.

References

[1] Badr Alkhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, 2024.

[2] Andrew Caines, Luca Benedetto, Shiva Taslimipoor, Christopher Davis, Yuan Gao, Øistein E Andersen, Zheng Yuan, Mark Elliott, Russell Moore, Christopher Bryant, et al. On the application of large language models for language teaching and assessment technology. In *LLM@ AIED*, 2023.

[3] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. Black-box access is insufficient for rigorous ai audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 2254–2272, 2024.

[4] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[5] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. Uncovering chatgpt's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1126–1132, 2023.

[6] MaryBeth Defrance, Maarten Buyl, and Tijl De Bie. Abcfair: an adaptable benchmark approach for comparing fairness methods. *Advances in Neural Information Processing Systems*, 37:40145–40163, 2024.

[7] Yashar Deldjoo. Understanding biases in chatgpt-based recommender systems: Provider fairness, temporal stability, and recency. *ACM Transactions on Recommender Systems*, 2024.

[8] Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[9] 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*, 50(3):1097–1179, 2024.

[10] Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. *Fundamentals of item response theory*, volume 2. Sage, 1991.

[11] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[12] Jianning Li, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. Chatgpt in healthcare: a taxonomy and systematic review. *Computer Methods and Programs in Biomedicine*, 245:108013, 2024.

[13] Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. Intersectional Stereotypes in Large Language Models: Dataset and Analysis. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8589–8597, Singapore, December 2023. Association for Computational Linguistics.

[14] Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C Treleaven, and Miguel Rodrigues Rodrigues. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, 2025.

[15] Huy Nghiem, John Prindle, Jieyu Zhao, and Hal Daumé Iii. "you gotta be a doctor, lin": An investigation of name-based bias of large language models in employment recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7268–7287, 2024.

[16] Ricardo Trainotti Rabonato and Lilian Berton. A systematic review of fairness in machine learning. *AI and Ethics*, 5(3):1943–1954, 2025.

[17] Mark D Reckase. Computerized adaptive testing: A good idea waiting for the right technology. 1988.

[18] Iain Weissburg, Sathvika Anand, Sharon Levy, and Haewon Jeong. LLMs are Biased Teachers: Evaluating LLM Bias in Personalized Education, February 2025.

[19] Kyra Wilson and Aylin Caliskan. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval, August 2024.

[20] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. Recommender systems in the era of large language models (llms). *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6889–6907, 2024.