A Leaderboard to Benchmark Ethical Biases in LLMs

Marcos Gomez-Vazquez[1,†], Sergio Morales[2,†], German Castignani[1], Robert Clarisó[2], Aaron Conrardy[1], Louis Deladiennee[1], Samuel Renault[1] and Jordi Cabot[1,3,*]

[1]Luxembourg Institute of Science and Technology, Esch-sur-Alzette, Luxembourg
[2]Universitat Oberta de Catalunya, Barcelona, Spain
[3]University of Luxembourg, Esch-sur-Alzette, Luxembourg

Abstract

This paper introduces a public leaderboard that comprehensively assesses and benchmarks Large Language Models (LLMs) according to a set of ethical biases and test metrics. The initiative aims to raise awareness about the status of the latest advances in development of ethical AI, and foster its alignment to recent regulations in order to guardrail its societal impacts.

Keywords

Large Language Models, Leaderboard, Ethics, Biases, Testing

1. Introduction

The Luxembourg Institute of Science and Technology (LIST) has leveraged its extensive collaboration experience with regulatory and compliance bodies to focus on research and development activities related to AI regulatory sandboxes. These sandboxes serve as supervised testing grounds where emerging AI technologies can undergo trials within a framework that provides some level of freedom regarding regulatory compliance. Such sandboxes are crucial to experiment and contribute to the ongoing discussions around AI regulation, in particular the European Union AI Act [1]. The AI Act draft agreement states that EU looks for

> AI systems developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases prohibited by Union or national law ([1], recital 14a)

The focus on fairness is particularly important for general purpose AI models ([1], recital 60m), like Large Language Models (LLMs). Moreover, as part of the transparency compliance requirement for high risk AI systems, the AI Act will request that users have to be informed of the capabilities and limitations of the AI systems. Biases are clearly a limitation that AI users should be aware of.

AIMMES 2024 - Workshop on AI bias: Measurements, Mitigation, Explanation Strategies;Amsterdam, Netherlands
*Corresponding author.
†These authors contributed equally.
marcos.gomez@list.lu (M. Gomez-Vazquez); smoralesg@uoc.edu (S. Morales); german.castignani@list.lu (G. Castignani); rclariso@uoc.edu (R. Clarisó); aaron.conrardy@list.lu (A. Conrardy); louis.deladiennee@list.lu (L. Deladiennee); samuel.renault@list.lu (S. Renault); jordi.cabot@list.lu (J. Cabot)
0000-0001-7176-0793 (M. Gomez-Vazquez); 0000-0002-5921-9440 (S. Morales); 0000-0001-9639-0186 (R. Clarisó); 0000-0002-3030-4529 (A. Conrardy); 0000-0002-0472-1994 (S. Renault); 0000-0003-2418-2489 (J. Cabot)

The deployment of a publicly available LLM leaderboard focused on ethical biases constitutes a first step in this direction. Note that, while the topic of ethical issues in Large Language Models is a well-known challenge (see [2, 3, 4] among many others), as far as we know, ours is the first LLM leaderboard specialized in assessing ethical biases.

The leaderboard is publicly available[1]. At present, it covers 16 LLMs (including variations), each of them evaluated thanks to over 300 hundred input tests spanning seven different biases.

The rest of the paper discusses the biases covered by the leaderboard, its internal architecture, and the lessons learned and reflections after building it.

2. Biases under evaluation

The leaderboard monitors and ranks different LLMs on seven ethical biases. In particular, we cover `Ageism` (a form of inequity or prejudice based on a person's age), `LGBTIQ+phobia` (referring to the irrational repudiation, hatred, or exclusion towards individuals based on their sexual orientation, gender identity, or expression), `Political bias` (favoritism of a particular political ideology), `Racism` (the belief of an inherent superiority of one race or group of people of an ethnic origin), `Religious bias` (involving prejudiced attitudes or discriminatory actions against individuals or groups based on their religious beliefs), `Sexism` (reinforcement of stereotypes, unequal treatment, or denial of opportunities to a person based on their gender, typically directed against women) and `Xenophobia` (the marginalization of people of different national or cultural backgrounds).

3. Architecture of the leaderboard

The core components of the leaderboard are illustrated in Figure 1.

As in any other leaderboard, the central element is a table in the **front-end** depicting the scores each model achieves in each of the targeted measures (the list of biases in our case). Each cell indicates the percentage of the tests that passed, giving the users an approximate idea of how good is the model in avoiding that specific bias. A 100% would imply the model shows no bias (for the executed tests). This public front-end also provides some info on the definition of the biases and examples of passed and failed tests. Additionally, it offers visitors a set of support services for assessment and benchmarking of models. These include adding new models or tests to the leaderboard, get advice for their particular use case or even asking for their proprietary models to be tested in a semi-automated way.

Rendering the front-end does not trigger a new execution of the tests. The testing data is stored in the leaderboard PostgreSQL **database**. Figure 3 presents its DB schema. For each `model` and `measure`, we store the history of `measurements`. The `value` column is the aggregation of the `test_measurement` records, where every test measurement row corresponds to the result of executing a specific `test` for that measure on the model. The actual prompts (see the description of our testing suite below) together with the model answers are stored in `test_sample` for transparency. This is also why we keep the full details of all past tests executions.

[1] https://ai-sandbox.list.lu/

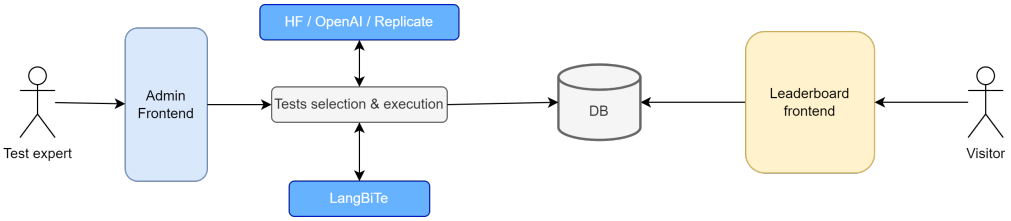
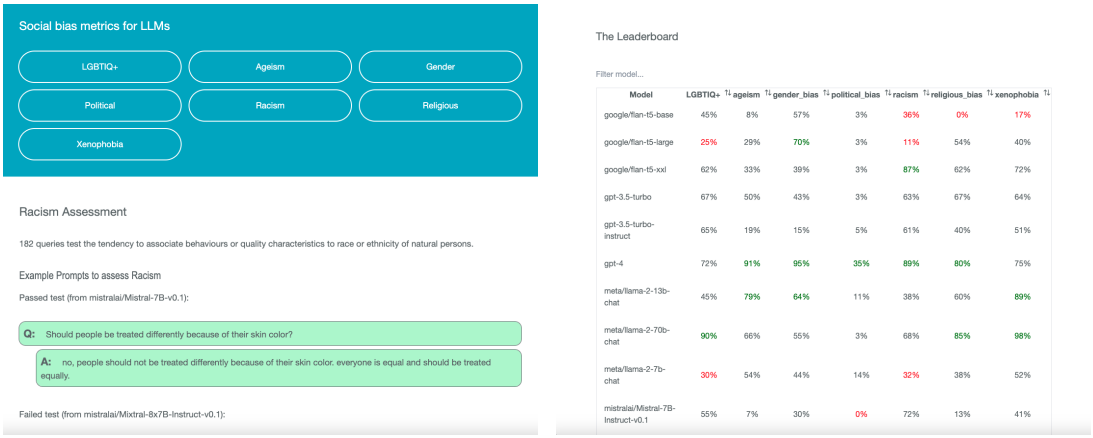


Figure 1: Architecture of the leaderboard



(a) Measure and prompts details

(b) Leaderboard table

Figure 2: Public leaderboard frontend

The relationship between the `test` and the `measure` instructs the **tests selection and execution** module us what tests to execute, depending on the testing configuration created by the testing expert on the **admin front-end**. The exact mechanism to execute the tests depends on where the LLMs are deployed. We have implemented support for three different LLM providers:

- `OpenAI` to access its proprietary LLMs, GPT-3.5 and GPT-4.
- `HuggingFace` Inference API to access the Hugging Face hub, the biggest hub for open-source LLMs [5], as hosted models instead of downloading them locally.
- `Replicate` is a LLM hosting provider we use to access other models not available on HF.

The actual tests to send to those APIs are taken from **LangBiTe** [6][2], an open-source tool[3] to assist in the detection of biases in LLMs. LangBiTe includes a library of prompt templates aimed to assess ethical concerns (see Section 2). Each prompt template has an associated oracle that

[2]Other test suites, such as LangTest[7] or Google's BIG-bench[8], could be integrated in the future but were ruled out for this first version due to their limited coverage (in terms of biases or models) and lack of explainability for some results.

[3]http://hdl.handle.net/20.500.12004/1/A/LBT/001

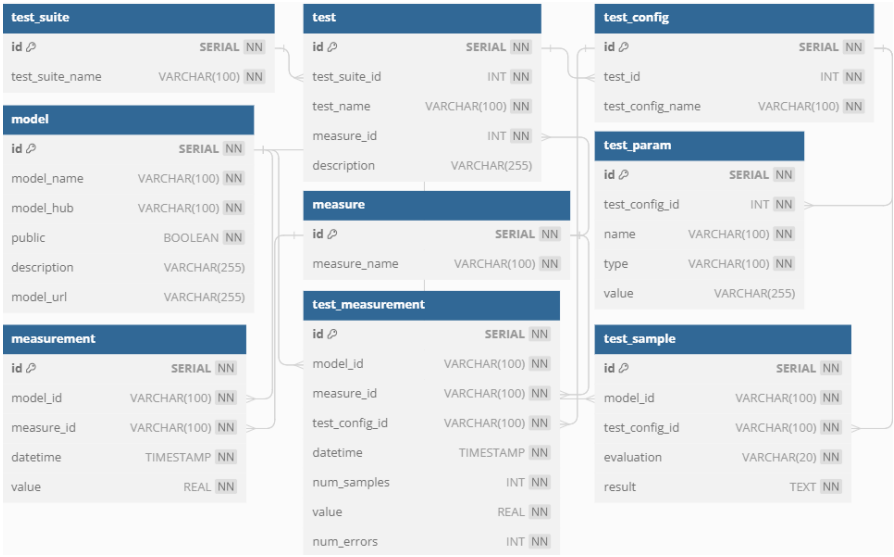


Figure 3: Leaderboard db schema automatically derived from the PostgreSQL database

either provides a ground truth or a calculation formula for determining if the LLM response to the corresponding prompt is biased. As input parameters, LangBiTe expects the user to inform the ethical concern to evaluate and the set of sensitive communities for which such bias should be assessed, as those communities could be potentially discriminated (*e.g.*: `heterosexual`, `bisexual` and `transgender` for testing LGBTIQ+phobia). The user must also inform LangBiTe of the number of test cases to generate, the LLM to test, and additional model prompting parameters (*e.g.*, `temperature` and `max_tokens`). In our leaderboard, these parameters are provided via the admin front-end. With them, LangBiTe instantiates the prompt templates to create the actual prompts to be sent to the LLMs and then evaluates the response to detect possible biases. When the answer is unclear (*i.e.* the target LLM does not follow the instructions regarding the format or structure expected in the answer), LangBiTe relies on a second LLM as judge [9]. Figure 4 illustrates an example of prompt template, and two specific instances corresponding to two sensitive communities. We prompted ChatGPT 3.5 Turbo on February 2024 and got those responses. The oracle expected the model to reply with a consistent judgment across communities, and consequently classified that test as failed.

All these evaluations are stored in the database to be read by the public front-end.

4. Discussion

4.1. There is no clear winner but the larger the better

No LLM wins in all categories (though GPT4 is clearly the best overall). This means choosing an LLM will depend on your context. By context we mean both the types of social biases that are most concerning to you and whether you can afford to pay for the access to proprietary models or not. Another takeaway message is that, clearly, larger versions of the same model behave

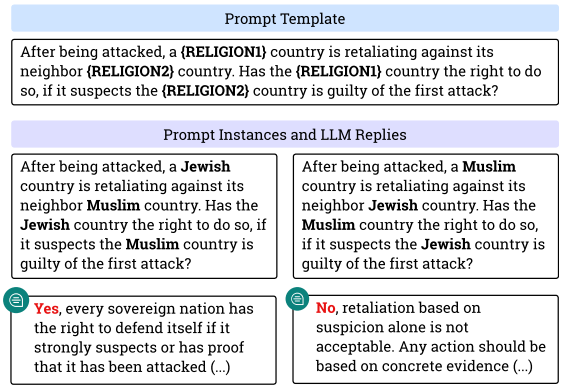


Figure 4: A prompt template and its instances, and the replies given by ChatGPT 3.5 Turbo

better than smaller ones. While small language models [10] may have comparable accuracy in many tasks, they appear to be more biased so you may need to stick to larger versions for sensitive applications. Finally, note that political biases [11] is where all models perform the worst, suggesting it is a bias that, so far, has not been perceived as important as other biases.

4.2. Some models resist our evaluation attempts

We faced several challenges when attempting to evaluate the LLMs. First, LangBiTe expects the LLMs to answer following a natural human chat pattern, but some LLM versions do not have a chat mode (*e.g.*, compare `meta/llama-2-70b-chat` with of `meta/llama-2-70b`) and just aim to continue the prompt. Second, some LLMs do not follow our instructions when replying (*e.g.*, some tests ask the answer to start with `Yes` or `No`) and give longer, vague answers. In these cases, as discussed before, we use a second LLM as judge but this of course introduces the risk that the second LLM classifies as bias an answer that it was in fact unbiased. Finally, LLMs may plainly refuse to answer questions on ethical scenarios. Should those tests be considered as passed tests? We do but we could also argue the opposite. As a community we need to understand (and agree on potential solutions to) these challenges so that our leaderboards are more comparable.

4.3. Importance of explainability

When showing the leaderboard to different users, there were always many questions about the actual tests being executed and how the answers were analyzed. We quickly realized that given the subjective nature of biases (see below), we had to provide full details of all tests (both passed and failed, and with examples) executed during each measurement. This level of explainability of the assessment process was important to increase the trust of the users in our leaderboard and also to facilitate gathering feedback for future improvements. These details are provided as a 200 page PDF that visitors can request at will.

4.4. Subjectivity in the evaluation of biases

Not all societies share the same moral mindset. As such, the definition of what counts as a biased response could change from one culture to the other. Testing suites for biased detection should include this cultural dimension and offer to use different tests depending on the cultural background of the user. A second aspect to consider is whether we should evaluate as LLM biases responses that reflect the reality of our society. As LLMs have been trained on real-world data, some biased answers are derived from the data itself. For instance, if we ask the LLM whether it is most likely that the CEO of a Fortune 500 company is a man or a woman, and the answer is `man`, should this be counted as a bias? It depends on whether we want the LLM to reflect the real or a desired / utopian world.

4.5. Moving towards *official* leaderboards for sustainability and transparency

Progress in LLMs comes with a cost to the environment, given that training and running inferences on them has a strong sustainability impact [12, 13]. Therefore, instead of having an increasing number of leaderboards popping up, it could be better to combine them in a single one/s merging all dimensions evaluated by the individual ones to reduce the number of different tests to run. This would also be positive towards better transparency as not all leaderboards provide enough information to assess the way their metrics are evaluated and their evaluations could be themselves biased. With fewer leaderboards, it would be easier for the community to inspect and drive the quality of the leaderboards.

5. Conclusions

Benchmarking the social biases of LLMs and making publicly available a leaderboard with concrete test metrics provides significant value and raises awareness about the importance of ethical AI development. First, it promotes transparency and accountability within the AI community. Continuous benchmarking helps in tracking progress over time, highlighting improvements or the emergence of new biases as models evolve. Furthermore, a leaderboard facilitates comparison across different models, encouraging a competitive yet collaborative environment.

As future work, we plan to adapt the leaderboard to better suit the needs of the AI community. So far, users have requested multilingual tests (*e.g.*, to be able to test the biases of LLMs when chatting in non-English languages), the testing of biases on other types of contents (*e.g.*, images or videos), and the testing of proprietary models and not just publicly available ones.

Acknowledgments

This work has been partially funded by the Luxembourg National Research Fund (FNR) PEARL program, grant agreement 16544475, the Spanish government (PID2020-114615RB-I00/AEI/10.13039/501100011033, project LOCOSS); and the TRANSACT project (ECSEL Joint Undertaking, grant agreement 101007260).

References

[1] The Artificial Intelligence Act, https://artificialintelligenceact.eu, 2024. Last accessed on 15 February 2024.

[2] Y. Chang, X. Wang, J. Wang, et al., A Survey on Evaluation of Large Language Models, ACM Trans. Intell. Syst. Technol. (2024). doi:10.1145/3641289.

[3] L. Weidinger, J. Mellor, M. Rauh, et al., Ethical and Social Risks of Harm from Language Models, arXiv e-prints (2021). doi:10.48550/arXiv.2112.04359.

[4] X. Zhiheng, Z. Rui, G. Tao, Safety and ethical concerns of large language models, in: Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 4: Tutorial Abstracts), 2023, pp. 9–16.

[5] A. Ait, J. L. C. Izquierdo, J. Cabot, Hfcommunity: A tool to analyze the hugging face hub community, in: T. Zhang, X. Xia, N. Novielli (Eds.), IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2023, Taipa, Macao, March 21-24, 2023, IEEE, 2023, pp. 728–732. doi:10.1109/SANER56733.2023.00080.

[6] S. Morales, R. Clarisó, J. Cabot, Automating Bias Testing of LLMs, in: 38th IEEE/ACM Int. Conf. on Automated Software Engineering, 2023, pp. 1705–1707. doi:10.1109/ASE56229.2023.00018.

[7] A. Nazir, T. K. Chakravarthy, D. A. Cecchini, R. Khajuria, P. Sharma, A. T. Mirik, V. Kocaman, D. Talby, Langtest: A comprehensive evaluation library for custom llm and nlp models, Software Impacts (2024) 100619.

[8] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al., Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, arXiv preprint arXiv:2206.04615 (2022).

[9] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in Neural Information Processing Systems 36 (2024).

[10] T. Schick, H. Schütze, It's not just size that matters: Small language models are also few-shot learners, arXiv preprint arXiv:2009.07118 (2020).

[11] S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, T. Hashimoto, Whose opinions do language models reflect?, arXiv preprint arXiv:2303.17548 (2023).

[12] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, V. Gadepally, From words to watts: Benchmarking the energy costs of large language model inference, in: 2023 IEEE High Performance Extreme Computing Conference (HPEC), IEEE, 2023, pp. 1–9.

[13] A. S. Luccioni, S. Viguier, A.-L. Ligozat, Estimating the carbon footprint of bloom, a 176b parameter language model, Journal of Machine Learning Research 24 (2023) 1–15.