Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Introducing MascuLead: the First Gender Bias Leaderboard

Fanny Ducel* Jeffrey André† Aurélie Névéol* Karën Fort†
* Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
† Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
*prenom.nom@lisn.fr, †prenom.nom@loria.fr

ABSTRACT

We present MascuLead, the first leaderboard centered on a gender bias detection task for inflected languages. We apply an existing framework on several Large Language Models and compare their performance. We also highlight the importance of including bias benchmarks in leaderboards, and question the very notion of leaderboards.

MOTS-CLÉS : biais, stéréotype, modèle de langue, genre, français.

KEYWORDS: bias, stereotype, large language model, gender, French.

1 On the Urge of Including Bias Benchmarks in Leaderboards

The field of Natural Language Processing (NLP) is experiencing exponential growth, leading to exponential development of Large Language Models (LLMs). The number of LLMs calls for robust evaluation methods, in order to identify state-of-the art models, but also for users to be able to pick the best-suited model for their needs. Leaderboards are a popular form of evaluation which consists in the publication of ranked LLMs, based on their performance on one or several tasks. For example, the *Open LLM Leaderboard* (Fourrier *et al.*, 2024) compares over 4,500 LLMs on 6 benchmarks, and has been liked by over 13,000 HuggingFace users[1]. A French version of this leaderboard is also available, covering about 30 LLMs on the same 6 translated benchmarks (Mohamad Alhajar, 2024).

Most leaderboards only include tasks such as question-answering, language understanding, or algorithmic reasoning. However, ethical aspects of LLMs are rarely taken into account, even though they can have a crucial impact on outputs and users. In this paper, we will focus on introducing stereotypical biases in NLP leaderboards. Numerous studies propose evaluation metrics for different types of stereotypical biases regarding gender (Choenni *et al.*, 2021), race (Hofmann *et al.*, 2024), socio-economic status (Cercas Curry *et al.*, 2024), etc. Even if this subfield of research focuses on English and US-centric stereotypes, efforts are carried out to address more diverse languages and socio-cultural contexts (Malik *et al.*, 2022; Fort *et al.*, 2024).

We believe that stereotypical biases should carry weight in leaderboards. Unbiased LLMs should be favored in comparison to biased and potentially harmful models. Further, including bias metrics in popular leaderboards would give more visibility to bias research. It could also "encourage researchers and engineers to pay more attention and direct more resources towards developing more exhaustive bias mitigation techniques and tackling more sources of biases" (Ducel *et al.*, 2024a).

Therefore, we introduce MascuLead, the first leaderboard evaluating gender biases for French LLMs.

[1] As of April 2025.

Actes de l'atelier Ethic and Alignment of (Large) Language Models 2025 (EALM), pages 12–19.

We use and extend the framework presented in Ducel *et al.* (2024b) to build our leaderboard. We also develop an online demonstrator to perform gender detection and bias evaluation, as well as update leaderboards[2]. The central task is cover letter generation, and gender biases are evaluated with two metrics: GenderGap and GenderShift (see Section 2). This framework presents advantages that make it a fit candidate to constitute a leaderboard task: it can be easily adapted to inflected languages and it does not rely on a corpus, which prevents overfitting[3]. We focus on French only and apply this framework on more LLMs than the original publication, generating 74,490 new cover letters with Mistral-7B-v0.3, Mistral-7B-Instruct-v0.3, CroissantLLMBase, CroissantLLMChat, gemma-2-2b, gemma-2-2b-it, Llama-3.2-3B, and Llama-3.2-3B-Instruct.

The additional models were chosen based on their popularity (they are among the most downloaded LLMs on HuggingFace as of April 2025) and the availability of both a base and Instruct version. The same two combinations of hyperparameters as the original article were selected[4]. For common models, we obtain the same results as the original paper, showing the reproducibility of the research.

2 Comparing LLMs with MascuLead

The resulting leaderboards are presented in Tables 1 to 3. The highest ranked LLMs are the ones that performed the best with the framework, and are supposedly the least biased according to the GenderGap (GG) and GenderShift (GS) metrics.

According to Ducel *et al.* (2024b), GG is the difference in proportions of masculine vs. feminine generated texts. Hence, the ideal GG is 0, i.e. there are as many masculine as feminine generated texts. A positive GG indicates a bias towards masculine markers, whereas a negative GG indicates a bias towards feminine markers. GS is the likelihood that texts are not consistent with the gender of the prompt (e.g., a model prompted with feminine markers generates a masculine cover letter). As we use either neutral or gendered prompts, we present separate leaderboards for each type of prompt.

Table 1 is an attempt at aggregating all three separate leaderboards. The ranking is based on the average of GG and GS. However, as GG uses a scale from -100 to 100, we convert the scores to absolute values, and separate them depending on the direction of the bias and of the type of prompts. Hence, the ideal values become 0, offering better compatibility with GS.

We still display the separate leaderboards (Table 2, 3), as the ranking changes and they illustrate various issues. Table 2 (left) unveils the models that are the more/less likely to favor a gender over the other, when the prompts do not contain gender information. Table 2 (right) highlights a similar phenomenon, but when using gendered prompts, which supposes that some texts do not respect the prompted gender. These two tables, based on GG, are related to representational harm as they imply stereotypical associations and the preference of one gender. Finally, Table 3 illustrates the likelihood that the gender of the prompt is overridden. This can be related to allocational harm, as users can be misgendered, and their request (a cover letter consistent with their gender) is not met.

All leaderboards highlight the important biases generated by the LLAMA models (including vigogne, a LLAMA model fine-tuned for French). *Mistral-7b-v0.3*, *xglm-2*, and *croissantbase* remain in the top 5 across all leaderboards. Table 1 allows for better generalization and readability of common points.

[2]The demonstrator is currently being deployed but the URL will be shared during the oral presentation.

[3]StereoSet (Nadeem *et al.*, 2021) and CrowsPairs (Nangia *et al.*, 2020) are widely used bias benchmarks. However, since they rely on corpora, models can be trained or fine-tuned on the material, an increasingly common phenomenon known as "data contamination" (Balloccu *et al.*, 2024; Deng *et al.*, 2024). In consequence, the bias metrics results can be misleading.

[4]top_p: 0.75 and top_k: 100 / top_p: 0.95 and top_k: 10 – except for Mistral, for which the first combination was replaced with temperature: 0.5 and top_p: 0.7 due to poor quality results with the initial combination

We can thus establish that all Instruct models appear more biased than their base counterparts. Further, the most biased models seem to be the most popular, but also best performing models according to traditional benchmarks (see Table 4 for a comparison with the *OpenLLMLeaderBoard* ranking).

Rank	Model	Avg (↓)	GG-masc-N	GG-fem-N	GG-masc-G	GG-fem-G	GS
1	*xglm-2*	13.64	1.08	/	7.05	/	32.79
2	mistral-7b-v0.3	17.87	0.71	/	/	7.73	45.18
3	croissantbase	24.98	/	8.15	9.07	/	57.71
4	*bloom-560m*	27.35	15.82	/	1.15	/	65.09
5	llama-3.2-3b	27.88	33.05	/	10.05	/	40.54
6	gemma-2-2b	30.27	23.7	/	10.39	/	56.71
7	*gpt2-fr*	31.66	12.81	/	21.81	/	60.35
8	*bloom-7b*	32.25	11.04	/	19.93	/	65.78
9	croissant-chat*	33.88	23.89	/	11.44	/	66.32
10	*bloom-3b*	36	18.95	/	17.23	/	71.82
11	mistral-7b-instruct-v0.3*	38.52	47.67	/	/	0.35	67.53
12	gemma-2-2b-it*	44.22	57.18	/	28.88	/	46.59
13	*vigogne-2-7b*	50.77	69.23	/	18.4	/	64.69
14	llama-3.2-3b-it*	58.14	65.57	/	25.47	/	83.37

Table 1: Global MascuLead. Italic: models used in the original paper. *: Instruct models.

3 Discussion – Leaderboards Flaws

MascuLead allows for a first evaluation and comparison of LLMs in terms of gender biases in French. However, this approach has several flaws, some of which are inherent to the notion of leaderboards. Leaderboards mostly focus on overall performance, which masks important biases, e.g. in relation to gender stereotypes. We argue that these grading systems should incorporate bias metrics in order to make biases, and the impact they have on performance and generations, more visible. By proposing an ethical reconfiguration of benchmarks, this contribution calls for the evolution of evaluation standards towards more responsible and equitable practices.

A closer look at the generated cover letters shows that data quality plays an important part in ranking models. Some models, such as *mistral-7b-v0.3*, appear to be among the least biased across all leaderboards, but exhibit major quality issues, e.g. irrelevant texts, gibberish[5], or the production of only a few words in French before switching to English[6]. Further analysis of the generations' quality could be conducted, in order to determine whether or not GS correlates with poor text consistency. If so, GS could be a relevant metrics to assess both gender bias and the general quality of generations.

These types of issues can question the very notion of leaderboards, since they only present aggregated scores without examples or error analyses. Moreover, aggregated scores often mix very different tasks, which can question the relevance and meaning of the scores. LLMs that have medium performance on all tasks may be ranked higher than some models that excel in specific tasks, fields or languages (including the task/field/language of interest for a specific user). Similarly, there does not seem to be a notion of weighting of scores, whereas some tasks may be more important than others. As shown by Raji *et al.* (2021) and Ethayarajh & Jurafsky (2020), the matter of general evaluation of LLMs and other NLP systems is at stake and should be further investigated. Minimally, stereotypical bias evaluation, as well as other types of ethical evaluations, should be included and carry more weight. We hope that MascuLead constitutes a first step towards bias inclusion in leaderboards, and that it will be expanded with other bias tasks, bias types, and more linguistic and cultural contexts.

[5] See an example in Appendix 3.

[6] A first naive approach based on language detection with GlotLID (Kargaran *et al.*, 2023) reveals that 48% of gendered texts and 40% of neutral texts generated with Mistral are labeled as English. See Tables 5 and 6 in Appendix for full results.

References

BALLOCCU S., SCHMIDTOVÁ P., LANGO M. & DUSEK O. (2024). Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In Y. GRAHAM & M. PURVER, Éds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 67–93, St. Julian's, Malta: Association for Computational Linguistics.

CERCAS CURRY A., ATTANASIO G., TALAT Z. & HOVY D. (2024). Classist tools: Social class correlates with performance in NLP. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 12643–12655, Bangkok, Thailand: Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.682.

CHOENNI R., SHUTOVA E. & VAN ROOIJ R. (2021). Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you? In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 1477–1491, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.111.

DENG C., ZHAO Y., TANG X., GERSTEIN M. & COHAN A. (2024). Investigating data contamination in modern benchmarks for large language models. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, p. 8706–8719, Mexico City, Mexico: Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.482.

DUCEL F., NÉVÉOL A. & FORT K. (2024a). Desiderata for Actionable Bias Research. In *New Perspectives on Bias and Discrimination in Language Technology*, Amsterdam, Netherlands. HAL : hal-04755691.

DUCEL F., NÉVÉOL A. & FORT K. (2024b). "You'll be a nurse, my son!" Automatically Assessing Gender Biases in Autoregressive Language Models in French and Italian. *Language Resources and Evaluation*. DOI : 10.1007/s10579-024-09780-6, HAL : hal-04803403.

ETHAYARAJH K. & JURAFSKY D. (2020). Utility is in the eye of the user: A critique of NLP leaderboards. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 4846–4853, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.393.

FORT K., ALONSO ALEMANY L., BENOTTI L., BEZANÇON J., BORG C., BORG M., CHEN Y., DUCEL F., DUPONT Y., IVETTA G., LI Z., MIESKES M., NAGUIB M., QIAN Y., RADAELLI M., SCHMEISSER-NIETO W. S., RAIMUNDO SCHULZ E., SACI T., SAIDI S., TORROBA MARCHANTE J., XIE S., ZANOTTO S. E. & NÉVÉOL A. (2024). Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 17764–17769, Torino, Italia: ELRA and ICCL.

FOURRIER C., HABIB N., LOZOVSKAYA A., SZAFER K. & WOLF T. (2024). Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard.

HOFMANN V., KALLURI P. R., JURAFSKY D. & KING S. (2024). Ai generates covertly racist decisions about people based on their dialect. *Nature*, **633**(8028), 147–154.

KARGARAN A. H., IMANI A., YVON F. & SCHUETZE H. (2023). GlotLID: Language identification for low-resource languages. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, p. 6155–6218, Singapore: Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.410.

MALIK V., DEV S., NISHI A., PENG N. & CHANG K.-W. (2022). Socially aware bias measurements for Hindi language representations. In M. CARPUAT, M.-C. DE MARNEFFE & I. V. MEZA RUIZ, Éds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 1041–1052, Seattle, United States: Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.76.

MOHAMAD ALHAJAR A. L. (2024). Open llm french leaderboard v0.2. https://huggingface.co/spaces/le-leadboard/OpenLLMFrenchLeaderboard.

NADEEM M., BETHKE A. & REDDY S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In C. ZONG, F. XIA, W. LI & R. NAVIGLI, Éds., *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)*, p. 5356–5371, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2021.acl-long.416.

NANGIA N., VANIA C., BHALERAO R. & BOWMAN S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In B. WEBBER, T. COHN, Y. HE & Y. LIU, Éds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1953–1967, Online: Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.154.

RAJI D., DENTON E., BENDER E. M., HANNA A. & PAULLADA A. (2021). Ai and the everything in the whole wide world benchmark. In J. VANSCHOREN & S. YEUNG, Éds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Appendix

Example of generated, gibberish text

"ça me semble être ce que je cherche. Je suis un homme de 30 ans, très sérieux et très motivé. Je suis à votre disposition pour plus de détails.

2017 Ford F-150

2017 Ford F-150 Interior Review

Note: This interior review was created when the 2017 Ford F-150 was new.

Scorecard

The 2017 Ford F-150 has a well-built cabin that's comfortable and spacious. The F-150's front seats are supportive and offer plenty of room for taller drivers. The rear seats are also comfortable and roomy, and the truck's cabin is quiet on the highway.

- "The F-150's cabin is a pleasant place to spend time, with comfortable"

Rank	Model	GenderGap (|x| = 0)
1	mistral-7b-v0.3	0.71
2	*xglm-2*	1.08
3	croissantbase	-8.15
4	*bloom-7b*	11.04
5	*gpt2-fr*	12.81
6	*bloom-560m*	15.82
7	*bloom-3b*	18.95
8	gemma-2-2b	23.70
9	croissant-chat*	23.89
10	llama-3.2-3b	33.05
11	mistral-7b-instruct-v0.3*	47.67
12	gemma-2-2b-it*	57.18
13	llama-3.2-3b-it*	65.57
14	*vigogne-2-7b*	69.23

Rank	Model	GenderGap (|x| = 0)
1	mistral-7b-instruct-v0.3*	-0.35
2	*bloom-560m*	1.15
3	*xglm-2*	7.05
4	mistral-7b-v0.3	-7.73
5	croissantbase	9.07
6	llama-3.2-3b	10.05
7	gemma-2-2b	10.39
8	croissant-chat*	11.44
9	*bloom-3b*	17.23
10	*vigogne-2-7b*	18.40
11	*bloom-7b*	19.93
12	*gpt2-fr*	21.81
13	llama-3.2-3b-it*	25.47
14	gemma-2-2b-it*	28.88

Table 2: MascuLead, on neutral (left) / gendered (right) prompts.

Rank	Model	GenderShift (↓)
1	*xglm-2*	32.79
2	llama-3.2-3b	40.54
3	mistral-7b-v0.3	45.18
4	gemma-2-2b-it*	46.59
5	gemma-2-2b	56.71
6	croissantbase	57.71
7	*gpt2-fr*	60.35
8	vigogne-2-7b	64.69
9	*bloom-560m*	65.09
10	*bloom-7b*	65.78
11	croissant-chat*	66.32
12	mistral-7b-instruct-v0.3*	67.53
13	*bloom-3b*	71.82
14	llama-3.2-3b-it*	83.37

Table 3: MascuLead, on gendered prompts, with GS as the key metrics.

MascuLead	Rank	Model	Avg (%)	Global Rank	Nb. downloads
14	1	Llama-3.2-3B-Instruct	24.2	1,768	11,592,453
11	2	Mistral-7B-Instruct-v0.3	19.23	2,799	11,556,197
12	3	gemma-2-2b-it	17.05	3,062	3,972,389
2	4	Mistral-7B-v0.3	14.58	3,390	6,797,298
6	5	gemma-2-2b	10.36	3,709	21,185,617
5	6	Llama-3.2-3B	8.7	3,809	3,740,053
10	7	bloom-3b	4.39	4,384	656,781
4	8	bloom-560m	3.51	4,525	21,939,835

Table 4: Scores of LLMs in the OpenLLMLeaderboard (English version), as of 04/29/2025. Global Rank is out of 4576 LLMs. Number of downloads is global ("all time"), from URLs such as this. *Note:* We use the English version of *OpenLLMLeaderboard* as the models of our experiment are not in the French version. Moreover, we hypothesize that most users would refer to the more popular, English version, even if to work on other languages

Rank	Model	EN (%)
1	mistral-7b-v0.3	40.16
2	llama-3.2-3b	15.32
3	mistral-7b-instruct-v0.3	3.65
4	gemma-2-2b-it	1.11
5	croissant-it	0.14
6	llama-3.2-3b-it	0.08
7	vigogne-2-7b	0.06
8	croissantbase	0.06
9	gpt2-fr	0.04
10	bloom-7b	0.02
11	gemma-2-2b	0.02
12	bloom-560m	0.00
13	bloom-3b	0.00
14	xglm-2	0.00

Table 5: Proportions of neutral generations that are labeled as English.

Rank	Model	EN (%)
1	mistral-7b-v0.3	48.71
2	llama-3.2-3b	13.14
3	mistral-7b-instruct-v0.3	8.98
4	gemma-2-2b-it	8.64
5	vigogne-2-7b	0.34
6	croissantbase	0.14
7	gemma-2-2b	0.12
8	gpt2-fr	0.04
9	llama-3.2-3b-it	0.04
10	bloom-560m	0.00
11	bloom-3b	0.00
12	xglm-2	0.00
13	bloom-7b	0.00
14	croissant-it	0.00

Table 6: Proportions of gendered generations that are labeled as English.