Laboratoire
Linguistique, Langues, Parole | LILPA | UR 1339
Université de Strasbourg

Analyzing use of masculine generics by LLMs in French

Enzo Doyen (LiLPa, University of Strasbourg)

1. Introduction

- Masculine generics (MG) in gender-marked languages (e.g., French, German, Dutch): use of the masculine form as a default/neutral form to refer to (a) mixed group of men/women or (b) people whose gender is unknown. Examples in French:

  (a) « Les **étudiants** sont partis. » (**Students** [masc.] left.)
  (b) « Un **athlète** doit s'entrainer régulièrement pour progresser. »
  (An **athlete** [masc.] needs to train regularly to progress.)

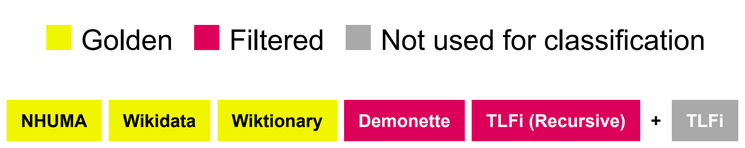
- Psycholinguistics studies show that **MG induce gender bias** and **amplify male-centric mental representations**[1–3]
- Gender bias widely studied in instruct-based LLMs[4], but never with generic instructions or in unconstrained contexts

2. Methodology

Focus on French, but applicable to any language with MG given human noun and instruction datasets

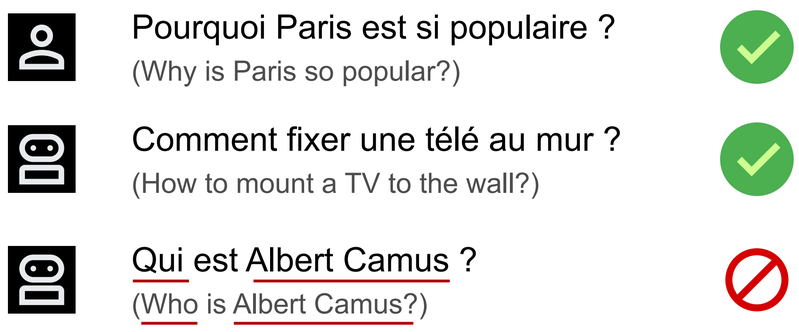
a. Create a French human noun (HN) dataset from available French lexical resources to detect occurrences of MG and evaluate the ratio of MG to HN uses

⊙ Filter with custom ML binary HN classification pipeline



b. Analyze MG use in 6 LLMs' outputs to generic instructions

⊙ Use human/AI-written generic instruction datasets and remove specific contexts with spaCy[5] (dependency parsing and NER)



⊙ Sample 10,000 instructions and send them to LLMs

⊙ Like instructions, filter responses to remove specific contexts

⊙ Validate HNs in outputs using GPT-4o mini, JSON-constrained

⊙ Compute score for each text; as well as mean (average bias per text) and overall (average bias per LLM) scores

$$\text{MScore}_i = \frac{\text{mg\_count}_i}{\text{hn\_count}_i}$$

Overall

$$\frac{\text{total\_mg}}{\text{total\_hn}}$$

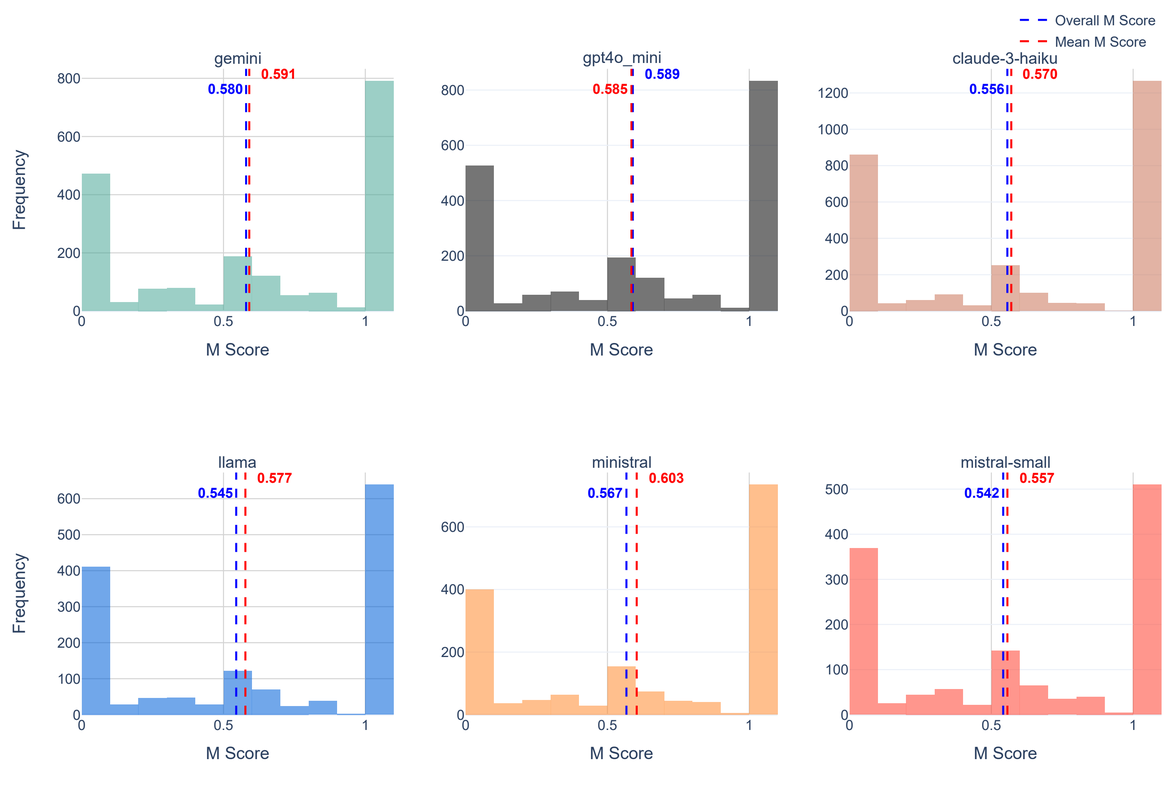
Mean

$$\frac{1}{\text{n}} \sum_{i=1}^{n} \text{MScore}_i$$

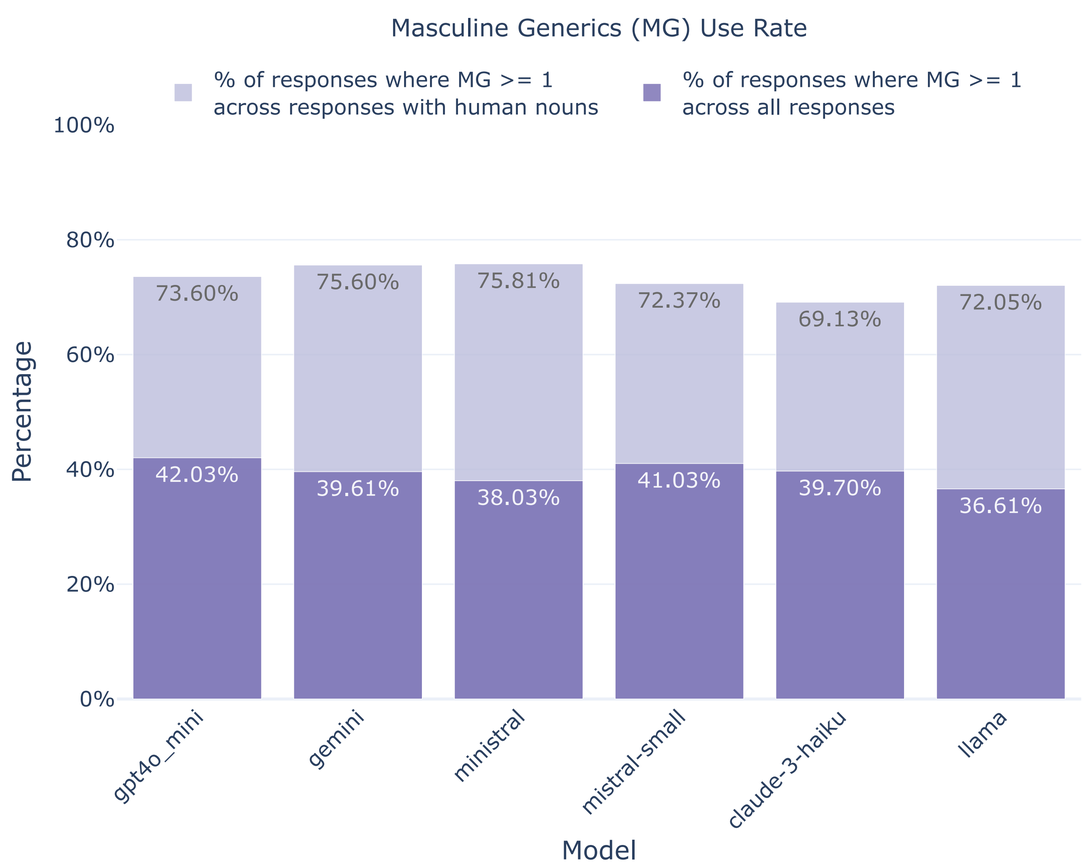
3. Results and Findings

- LLMs use MG in ≈39.5% of all their responses on average (≈73.1% of responses with HNs)
- GPT-4o mini and Ministral generally the most biased models
- Llama 3 8B, Claude 3 Haiku and Mistral Small 3 generally the least biased models
- LLMs reluctant to using gender-fair language (GFL) spontaneously, Llama 3 8B being the model with highest GFL use (see preprint for details)

MScore



MG Use Rate



4. Takeaways

- LLMs largely exhibit MG bias when generating responses to generic, contextually unconstrained instructions
- Fairness in language should be attentively considered when training LLMs in heavily gender-marked languages

References

[1] Braun, F., Sczesny, S., & Stahlberg, D. (2005). Cognitive Effects of Masculine Generics in German: An Overview of Empirical Findings. *Communications*, *30*(1), 1–21.
[2] Gygax, P., Gabriel, U., Lévy, A., Pool, E., Grivel, M., & Pedrazzini, E. (2012). The Masculine Form and Its Competing Interpretations in French: When Linking Grammatically Masculine Role Names to Female Referents Is Difficult. *Journal of Cognitive Psychology*, *24*(4), 395–408.
[3] Rothermund, P., & Strack, F. (2024). Reminding May Not Be Enough: Overcoming the Male Dominance of the Generic Masculine. *Journal of Language and Social Psychology*, *43*(4), 468–485.
[4] Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., … Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. arXiv.
[5] Montani, I., Honnibal, M., Boyd, A., Landeghem, S. V., & Peters, H. (2024). spaCy: Industrial-strength Natural Language Processing in Python.

Read our full preprint for more details:
Doyen, E. & Todirascu, A. (2025). *Man Made Language Models? Evaluating LLMs' Perpetuation of Masculine Generics Bias*. arXiv: 2502.10577.

ALPS 2025