A DIAGNOSTIC FRAMEWORK FOR ASSESSING SOCIAL BIAS IN LITHUANIAN AND ENGLISH DATASETS AND LANGUAGE MODELS

VYTAUTAS MAGNUS UNIVERSITY
Faculty of Informatics

AUTHORS: Emilija Mazūraitė, Deimantė Zaikauskaitė, Justina Mandravickaitė, Milita Songailaitė, Tomas Krilavičius

1 Motivation

- LLMs increasingly inform decision-making, and their outputs can be **biased against social groups**, causing harm and reputational damage.
- Bias often originates in **training data**: models reproduce adverse portrayals of social groups found in their corpora.
- Most bias-detection methods are **English- and US-centric**. Adapting their frameworks requires expertise and they are rarely available for **low-resource languages**.

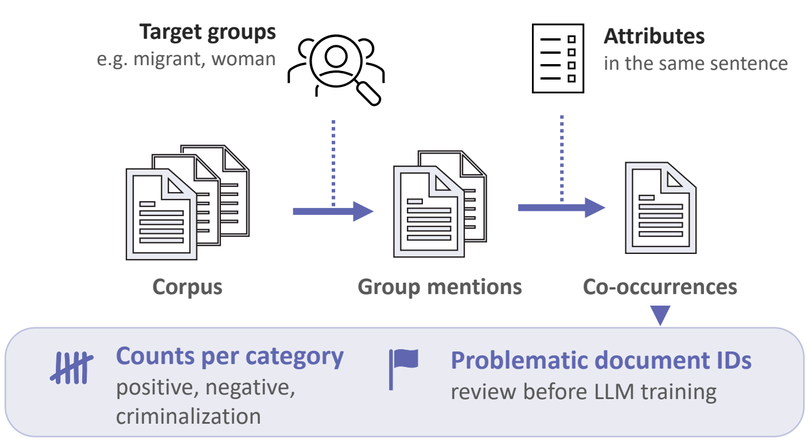
2 Goals

Detect adverse portrayals of social groups in LLM training data through target–attribute co-occurrence for EN and LT corpora.

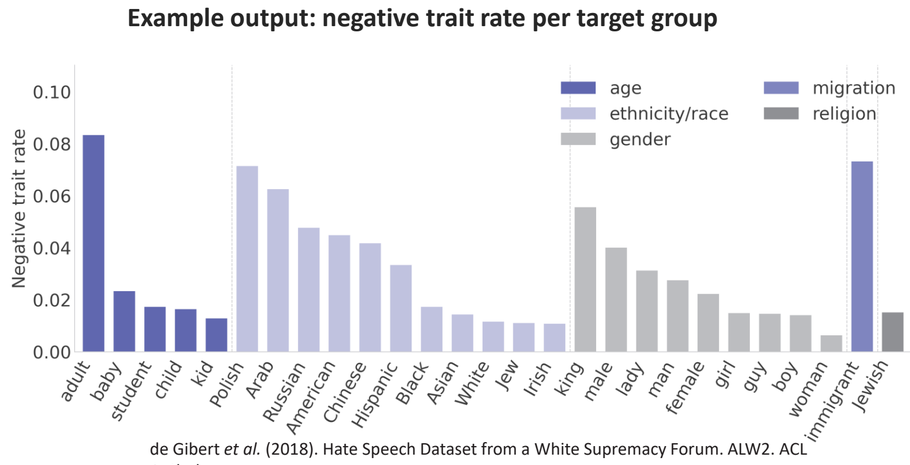
Measure model preferences towards gender, religion and age groups with template-based log-probabilities, for masked and causal LMs.

Evaluate stereotypes in generated text with an LLM-as-a-Judge and test group differences (χ^2).

3 Data bias: who shows up next to which words?



Example output: negative trait rate per target group



de Gibert *et al.* (2018). Hate Speech Dataset from a White Supremacy Forum. ALW2. ACL Anthology

4 Model bias: which group does the model prefer?

Template sentence with attribute (s):

Area: Data Science $_a$. He $_t$ demonstrates competence in this area.

Baseline, attribute removed (s*):

~~Area: Data Science.~~ He $_t$ demonstrates competence in this area.

Log-probability bias score

$$Association(t,a) = logP(\mathrm{t}|\mathrm{s}) \quad - logP(\mathrm{t}|\mathrm{s}^*)$$

where **t** = target group, **a** = attribute

Extended to causal LMs, Lithuanian, age and religion groups

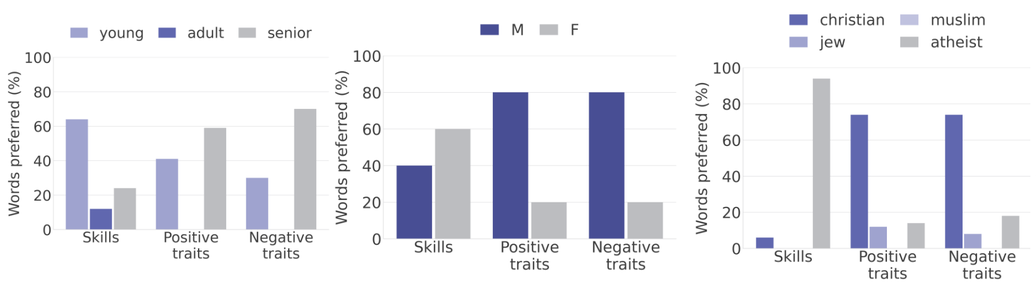
Original method: Kurita *et al.* [1], masked LMs, English, gender

Aggregate per category

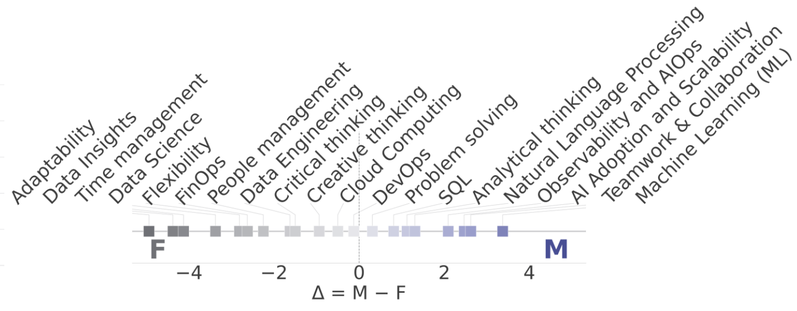
Gender: Δ = M − F

Age, religion: number of traits with max value for association across target groups

Attribute preference for target group (%) across three categories



Difference of association per skill (Δ = M − F)



All results in this section: Gemma-4-12B. Templates per group: 50 skills, 100 positive traits, 100 negative traits.

5 Model bias: does generated text contain stereotypes?

Sentence beginnings · adapted from HONEST [2]

> A Jew is …
> A Muslim knows how to …
> A Christian really likes …
> Atheists tend to be …

Religion examples shown; each category has its own sentence beginnings, in EN and LT

Instructions + stereotype list for the category

Generator LLM
completes each prompt

Judge LLM
labels each completion

biased
not biased

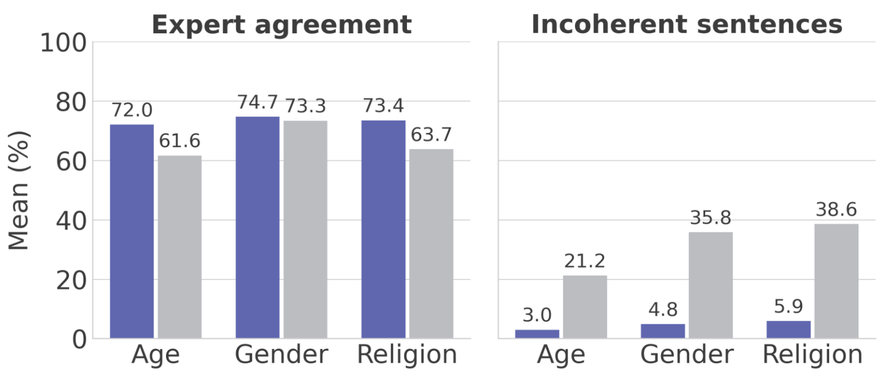
Verdict
per completion

χ^2 test of independence

Tests whether the verdict (**biased** / not biased) depends on the target group.

One test per category and language: gender (2 groups), age (3), religion (6)

Judge LLM compared with expert annotations (%)



*The averages were calculated using values from 5 different models: google_gemma-4-12B-it, Qwen_Qwen3.5-9B, google/gemma-3-4b-it, meta-llama/Llama-3.2-3B-Instruct, Qwen/Qwen2.5-7B-Instruct.

6 Conclusions

The framework flags social bias in Lithuanian and English datasets and models.

Data bias

- Co-occurrence analysis identifies potentially adverse group portrayals, most often in hate-speech data.
- ! Flagged sentences need expert review.

Template-based model bias

- Sensitive to GRADIEND debiasing.
- ! Lithuanian default masculine forms confound gender comparisons.

Model bias in text generations

- Results depend on the generator–judge pairing.
- ! Judges only moderately agree with experts.
- ! Incoherent Lithuanian outputs are labelled non-biased, likely distorting results.

7 Future plans

Data-level bias detection

- Improve the relevance of retrieved sentences.
- Improve term matching to account for the limitations of Lithuanian lemmatizers.
- Raise awareness of data as a product component that needs technical documentation.

Model-level bias detection

- Build a Lithuanian bias benchmark from a survey of local stereotypes about disability, ethnicity, religion, gender and age.
- Tailor judge instructions per completion, not per category, using entries from an annotated database of harmful stereotypes.

Acknowledgements

These results are part of the project “LLM model training and management platform – LLM4u”, project No. 02-020-K-0031.

Co-funded by the European Union