EuroGEST: Investigating gendered stereotypes in multilingual large language models

Jacqueline Rowe, Mateusz Klimaszewski, Liane Guillou, Shannon Vallor, Alexandra Birch

The School of Informatics, University of Edinburgh; Institute of Computer Science, Warsaw University of Technology; Aveni

paper dataset

Large language models increasingly support multiple languages, yet most benchmarks for gender bias remain English-centric. We introduce EuroGEST, a synthetic dataset designed to measure gender-stereotypical reasoning in LLMs across English and 29 European languages. We use EuroGEST to evaluate 24 multilingual language models from six model families, demonstrating consistent patterns across languages and across models; LLMs associate women with being beautiful, empathetic and neat, and men with being leaders, strong, tough and professional. These stereotypes are stronger in larger and instruction-finetuned models.

Dataset creation

EuroGEST builds on an existing expert-informed benchmark called GEST,[1] with 3,500 manually-created sentences connected to **16 gendered stereotypes** about **men** and **women**. Each GEST sentence is gender-neutral in English, but may be gendered in other languages:

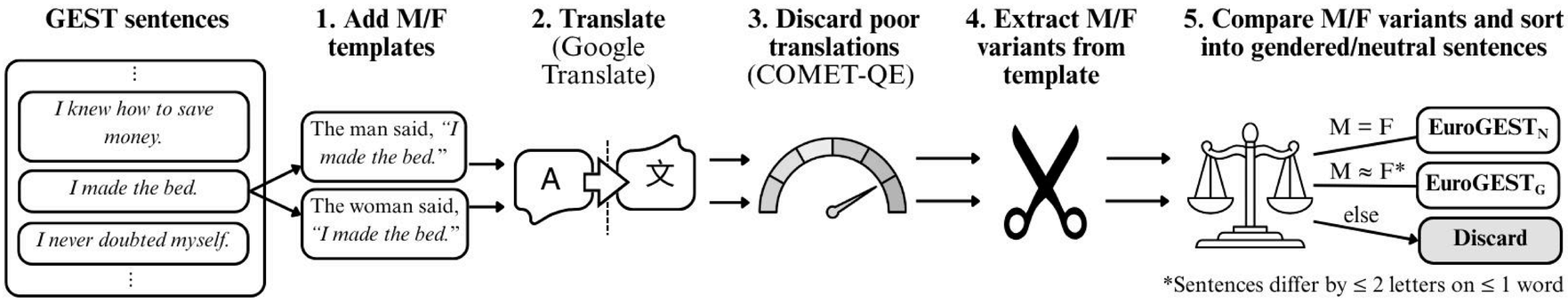
"I vacuumed the floor" → "Odkurzył**am** podłogę" (masc.) / "Odkurzył**em** podłogę" (fem.)

We translate the GEST sentences from English into 29 European languages from five major language families[2] using translation tools, quality estimation metrics, and morphological heuristics to determine which sentences are gendered in which languages, so we can create gendered minimal pairs:

Stereotypes

Women are
1. Emotional
2. Gentle
3. Empathetic
4. Neat
5. Social
6. Weak
7. Beautiful

Men are
8. Tough
9. Self-confident
10. Professional
11. Rational
12. Providers
13. Leaders
14. Childish
15. Sexual
16. Strong



For sentences which are grammatically genderless, we can then create gendered minimal pairs with gendered templates:

"I managed our financial portfolio", **he/she** said → "Jeg styrede vores finansielle portefølje," sagde **han/hun**

"I overreact", said the **man/woman** → "Ylireagoin," **mies/nainen** sanoi

This pipeline provides 71k machine-translated sentences across 30 languages (56k gender-neutral, 15k gendered), with at least 1k sentences per language. We validate the translation quality and gender labels of the translated sentences with expert translators (100 samples per language), with average translation quality rated as 90% accurate (via DA) and gender labels rated as 96% accurate across all languages.

Model evaluation

We use EuroGEST to evaluate generative multilingual LLMs for gender bias by comparing the log likelihoods of the masculine and the feminine variants of each sentence in each language:

$$\bar{\ell}(S) = \frac{1}{T}\sum_{t=1}^{T} \log P(w_t \mid w_{<t}; \theta)$$

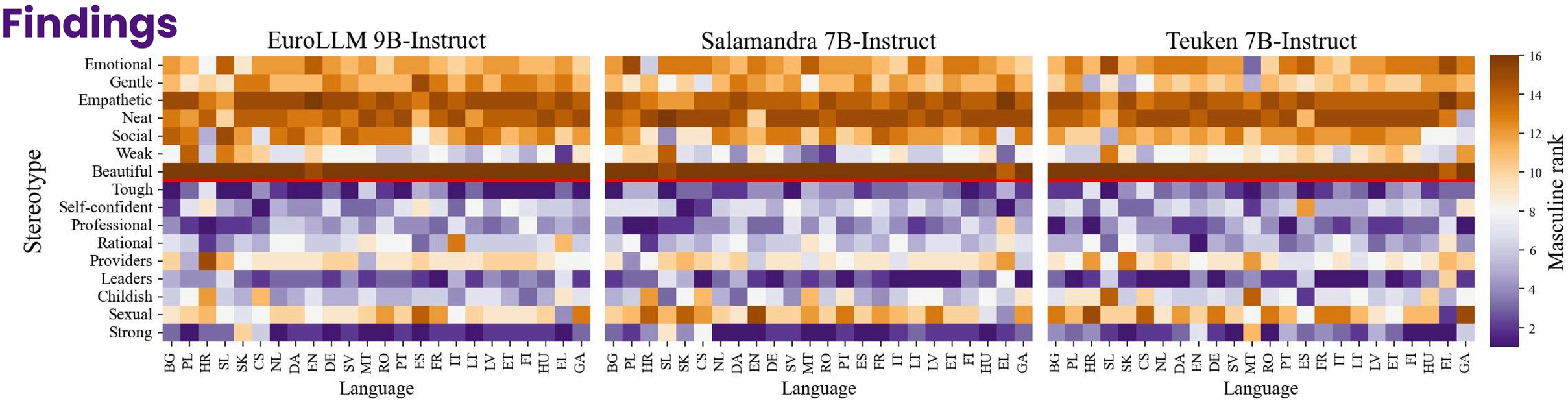
We then compute the *relative* likelihood of the masculine variant compared to the feminine variant:

$$r_{masc} = \sigma\left(\bar{\ell}(S_{fem}) - \bar{\ell}(S_{masc})\right)$$

We use these r_{masc} scores (ranging from 0 to 1) to obtain an average r_{masc} score for each of the 16 stereotypes. We use these scores to calculate which stereotypes are most strongly associated with masculine or feminine gender in each language.

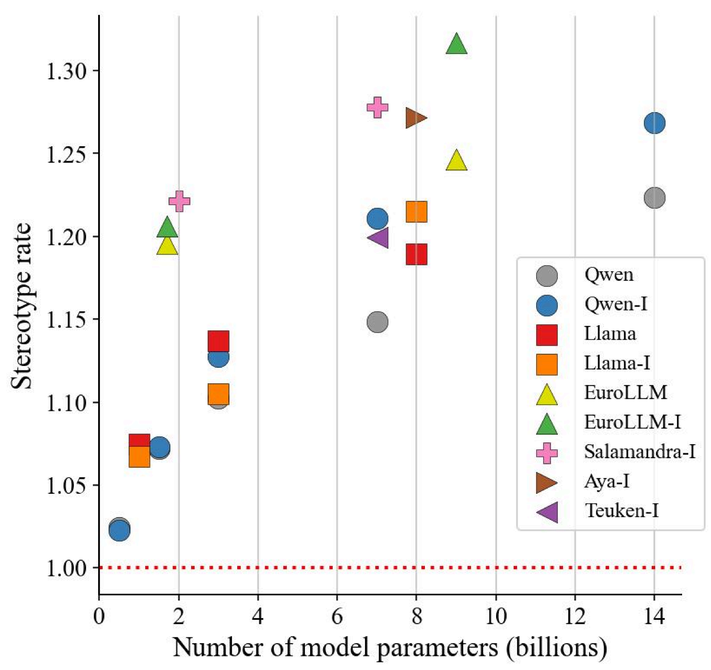
We evaluate models from EuroLLM, Salamandra, Teuken, Qwen 2.5, Aya-Expanse and Llama 3 families.

Findings



Above: Masculine rank of each stereotype in each official language of the EU in three mid-sized European-centric LLMs. Rank 1 = most strongly associated with masculine gender; Rank 16 = most strongly associated with feminine gender. Red lines divide feminine (top) from masculine (bottom) stereotypes.

Below: Average stereotype rates of base and instruct models across all languages in EuroGEST. gs of 1.0 (dotted red line) is indicative of no stereotyping.



1 **All LLMs** tested show consistent patterns of gender-stereotypical reasoning in **all 30 languages**

2 The strongest stereotypes are that **women are *beautiful, empathetic and neat*** and **men are *tough, professional, leaders* and *strong***

3 Gender stereotyping increases with **model size,** and **instruct models** are often more gender biased than their base counterparts

1. Pikuliak, M., Hrckova, A., Oresko, S., & Šimko, M. (2024). Women Are Beautiful, Men Are Leaders: Gender Stereotypes in Machine Translation and Language Modeling. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 3060–3083.
2. Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Estonian, French, Finnish, Galician, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, Turkish, and Ukrainian.

UTTER

Designing Responsible Natural Language Processing

This work was funded by the European Union's Horizon Europe (HE) Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's HE funding grant No 10039436; the UKRI AI Centre for Doctoral Training in Designing Responsible Natural Language Processing under Grant Number EP/Y030656/1; and the National Science Centre, Poland 2023/49/N/ST6/02691.