EuroGEST: Investigating gender stereotypes in multilingual language models

Anonymous ACL submission

Abstract

Large language models increasingly support multiple languages, yet most benchmarks for gender bias remain English-centric. We introduce EuroGEST, a dataset designed to measure gender-stereotypical reasoning in LLMs across English and 29 European languages. EuroGEST builds on an existing expert-informed benchmark covering 16 gender stereotypes, expanded in this work using translation tools, quality estimation metrics, and morphological heuristics. Human evaluations confirm that our data generation method results in high accuracy of both translations and gender labels across languages. We use EuroGEST to evaluate 24 multilingual language models from six model families, demonstrating that the strongest stereotypes in all models across all languages are that women are *beautiful, empathetic* and *neat* and men are *leaders*, *strong, tough* and *professional*. We also show that larger models encode gendered stereotypes more strongly and that instruction finetuning does not consistently reduce gendered stereotypes. Our work highlights the need for more multilingual studies of fairness in LLMs and offers scalable methods and resources to audit gender bias across languages.

1 Introduction

Large language models (LLMs) encode social biases (Barikeri et al., 2021; Gallegos et al., 2024; Gupta et al., 2024; Gemini Team et al., 2024; Parrish et al., 2022; Sanh et al., 2021; Smith et al., 2022). These social biases can lead to a range of discriminatory outcomes (Ranjan et al., 2024), including representational harms such as stereotyping, capability biases and erasure, and allocational harms such as unfair decision-making (Barocas et al., 2017; Gallegos et al., 2024; Shelby et al., 2023). Bias benchmarks can serve as useful tools for evaluating and calling attention to systemic LLM biases that may cause social harms in certain contexts, provided that the motivations, values and norms embedded in the benchmark design are clearly articulated (Blodgett et al., 2020; Goldfarb-Tarrant et al., 2023).

Most existing bias benchmarks serve only a few high-resource languages (Blodgett et al., 2020; Röttger et al., 2024), and there are few studies exploring how social biases in LLMs vary by language, or how to design benchmarks that translate well across languages. As such, developers of widely-used multilingual LLMs do not evaluate for bias in all supported languages (see, for example, Grattafiori et al. (2024); Martins et al. (2024); Team et al. (2022); Üstün et al. (2024)), meaning that little is known about whether existing bias mitigation techniques effectively prevent discriminatory across different languages.

In this work, we focus on measuring gendered stereotypes in multilingual generative LLMs. While LLMs exhibit a wide range of social biases, gender is a salient and universally encoded dimension of identity, and gender roles and stereotypes are systematically embedded in language usage across cultures. Many languages encode gender at the level of morphology and pronoun systems, making multilingual investigation of gender biases in LLMs challenging. To address this, we introduce EuroGEST,[1] a new gender bias benchmark dataset that adapts and extends an existing open-source gender bias benchmark (Pikuliak et al., 2024) to cover 29 additional European languages from five main language families.[2] We focus on European languages because they are relatively highly resourced, facilitating automatic scaling of benchmark data, and because this group includes lan-

[1] Available at [github repository anonymised for review] under an Apache 2.0 license.

[2] Slavic: Bulgarian, Croatian, Czech, Polish, Russian, Slovak, Slovenian, Ukrainian. Germanic: Danish, Dutch, English, German, Norwegian, Swedish. Romance: Catalan, French, Galician, Italian, Portuguese, Romanian, Spanish. Baltic: Latvian, Lithuanian. Uralic: Estonian, Finnish, Hungarian. Other: Greek, Irish, Maltese and Turkish.

guages with gendered morphology as well as languages which are grammatically gender-neutral. Cultural and socio-economic parallels across Europe also make expressions of gender bias more comparable within this region than between European and non-European contexts.

Our main contributions are as follows:

- We introduce EuroGEST, a novel dataset of 71,000 sentences linked to 16 gendered stereotypes across 30 European languages;
- We develop an automated pipeline that combining linguistic expertise, machine translation and quality estimation to efficiently and cheaply generate accurately-labelled gendered minimal pairs;
- We provide cross-lingual evidence that multilingual language models systematically amplify similar gendered stereotypes across diverse European languages;
- We show that larger language models encode these stereotypes more strongly, and that instruction finetuning does not effectively mitigate these gendered biases.

We hope that our methodology, dataset and results will spur more in-depth and fine-grained investigations of how LLMs manifest social biases in different linguistic and cultural contexts.

2 Related work

Previous investigations into how gender biases surface in NLP tools and LLMs in particular have covered a wide range of topics, tasks, intersectional identities and empirical methods (Bartl et al., 2025; Blodgett et al., 2020; Gallegos et al., 2024; Stanczak and Augenstein, 2021). Gender is expressed and performed in language in complex ways, so no single method or approach will provide a holistic picture of 'gender biasedness' in an LLM in different languages and cultures. Here we summarise existing techniques and highlight gaps with regard to multilingual gender bias detection.

Much work has focused on measuring extrinsic gender biases exhibited by LLMs. The widely-used BBQ dataset (Parrish et al., 2022) fills 25 question templates with indicators for different social demographics (including gender), measuring bias in terms of whether the LLM's responses to the questions correspond to stereotypes or not. Similarly, Gupta et al. (2024) create slot-filled templates from existing NLU benchmarks, testing model responses with different proper names associated with different demographic groups to investigate whether the LLM exhibits bias in performance on the task. Tamkin et al. (2023) create prompt templates for investigating bias in realistic decisionmaking scenarios spanning finance, business, law and education, and for text generation Kirk et al. (2021), Lucy and Bamman (2021) and Wan et al. (2023) explore gender biases displayed by LLMs in sentence completion, storywriting, and reference letter drafting tasks. Multilingual extrinsic bias evaluations have typically focused on exploring whether translations from gender-neutral into gendered languages follow stereotypical biases (Savoldi et al., 2021; Stanovsky et al., 2019; Bentivogli et al., 2020).

While these studies have a helpful focus on how gendered harms might arise as LLMs are utilised in practice, their data generation and analysis methods are difficult to scale and compare across languages, and most of their datasets are available only in English. Other work focuses on investigating intrinsic bias in LLMs' internal representations rather than their outputs. For example, minimal gendered pairs from the Winogender (Rudinger et al., 2018) and Winobias (Zhao et al., 2018) coreference bias datasets can be passed to LLMs as prompts, to compare whether the LLM assigns greater likelihoods for stereotypically-gendered sentences (Glaese et al., 2022). Nangia et al. (2020) and Pikuliak et al. (2024) take the same approach using gendered minimal pairs from the CrowS-Pairs and GEST datasets respectively, and Barikeri et al. (2021) compare the perplexity of stereotypical and anti-stereotypical Reddit comments.

These intrinsic bias evaluation methods are distant from practical application and may not be relevant for certain use cases, but they can still highlight strongly encoded biases that may need further testing and mitigation in certain contexts. They can also be scaled across many languages to provide a more multilingual picture of how LLMs encode gender biases. For example, Pikuliak et al. (2024) utilise gendered minimal pairs in English and nine Slavic languages to assess gender bias in masked and generative language models, and Mitchell et al. (2025) measure bias in 16 different languages by measuring token likelihoods on manually-curated and translated gendered minimal pairs. Manual curation of prompts by local and native speakers of each language context, as demonstrated in Mitchell et al. (2025), produces multi-

lingual benchmark data which is well-adapted to linguistic and cultural differences in how bias is expressed (Borah et al., 2025; Dev et al., 2023; Myung et al., 2024). However, such approaches are highly resource-intensive, and there is an interim need for more rapidly scaleable methods to expand bias benchmarks across a greater range of languages, both to investigate possible gendered representational harms and to test the efficacy of different gender bias mitigation techniques across languages.

3 Method

3.1 The Original GEST dataset

We chose to adapt and extend the GEST dataset (Pikuliak et al., 2024) because its construction is informed by gender expertise and it includes a large number of sentences (3,500) explicitly focused on gendered stereotypes. Furthermore, the GEST dataset utilises heuristics for morphological gender detection in Slavic languages which are applicable to other European languages which express morphological gender on nouns, adjectives and verbs with word-final suffixes. Finally, the authors demonstrate how GEST data can be used to measure gender bias in both richly-gendered languages and gender-neutral languages, making it suitable for the range of languages in our set.

To create GEST, Pikuliak et al. (2024) worked with gender experts to identify 16 common gendered stereotypes about men and women (listed in Appendix A). They then generated first-person English sentences associated with each stereotype, and created gendered minimal pairs of each sentence in each language. For English, they achieve this by wrapping the gender-neutral sentences in gendered templates (e.g, *'"I am emotional", he/she said'*); for Slavic languages, they use the morphologically masculine and feminine variants of the sentence (e.g. *"Som emotívny"* (masculine) or *"Som emotívna"* (feminine), in Slovak). They use these gendered minimal pairs to explore gender bias in both translation and text generation tasks.

3.2 Dataset Expansion

Of the 30 European languages we consider, 20 express gender on adjectives, nouns or verbs; 6 express morphological gender only on pronouns (including English); and 4 are not morphologically gendered at all (see Appendix B.1). The 20 gendered languages do not express gender on all of the GEST sentences; for example, sentence *'I started my own company when I was 18'* is gender-neutral in Italian (*'Ho fondato la mia azienda quando avevo 18 anni'*) while the sentence 'I gave up easily without a fight' is gendered (*'Mi sono arres**o/a** facilmente, senza combattere'*). We need to know whether each GEST sentence is gendered in a given language in order to create the appropriate gendered minimal pair.

To account for this variation, we wrap each English GEST sentence S in masculine and feminine templates (*'The man/woman said "S"'*) before translating both variants into all 20 gender-sensitive languages using the Google Translate API. We use the COMET Quality Estimation (QE) tool (Rei et al., 2020) to discard any translated sentences with QE score < 0.85, selecting the highest possible QE threshold that retains at least 1,000 sentences per language at the end of dataset creation (see Figure 2). We extract the masculine and feminine translations of each original GEST sentence and compare them. Identical pairs are assumed to be gender neutral sentences, and are added to the EuroGEST-neutral dataset (EuroGEST$_N$). Following Pikuliak et al. (2024), translations which differ by up to two letters on a single word are assumed to be a minimal gendered pair and are added to the EuroGEST-gendered dataset (EuroGEST$_G$). Sentence pairs differing by more than two letters on one word are discarded (see Figure 1 for an illustration). For the nine languages that, like English, lack morphological gender in first-person sentences, we simply translate the original GEST sentence, filter by COMET quality, and add the remaining translations to EuroGEST$_N$.

With this method, we obtain 14,538 sentences in EuroGEST$_N$ and 56,497 sentences in EuroGEST$_G$, with between 1121 (Maltese) and 3361 (Swedish) sentences per language (Figure 2). At the quality-estimation stage, the highest numbers of sentences are discarded from low-resource languages like Catalan, Irish, Galician and Maltese (see Appendix B.2), likely due to poor performance by Google Translate and/or COMET QE.

3.3 Dataset validation

To check the reliability of our automatically translated and labelled data, we manually evaluate 100 sample sentences per language. Following Kocmi et al. (2024), we ask expert translators to directly assess translation quality of each sentence on a scale of 0 to 100, providing boundaries to guide

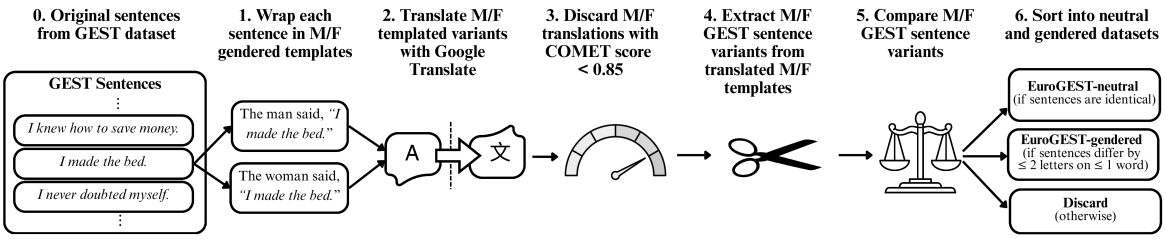

Figure 1: System for translating English GEST sentences into gendered target languages and sorting into gendered and gender-neutral pairs.

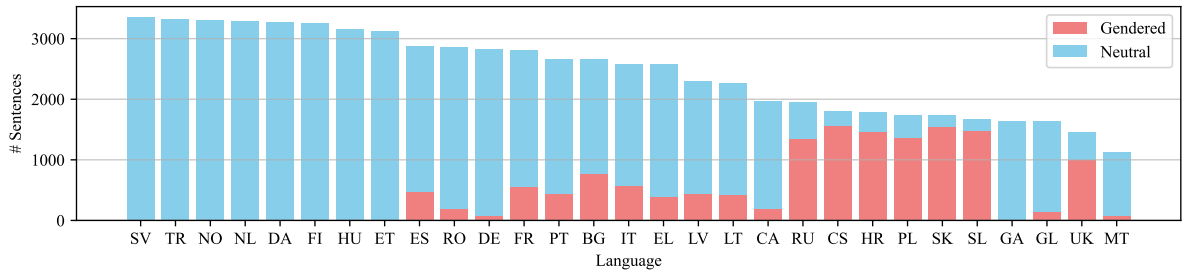

Figure 2: Number of sentences in EuroGEST-gendered and EuroGEST-neutral datasets by language

judgements (Appendix C.1). The average translation quality was rated as 90.1/100, and each language apart from Maltese had an average translation quality of over 80.0/100 (Appendix C.2). The translators also rated the gendered labels (neutral, masculine or feminine) of the sample sentences as accurate in 95.9% of cases across all 30 languages, and 94.5% across only the gendered languages (for which gender labeling is more difficult). Maltese showed the lowest gender label accuracy at 85%, and due to these poor validation results we excluded Maltese data from our experiments.

3.4 Model Evaluation

We use EuroGEST to evaluate generative multilingual LLMs for gender bias by exploring whether the LLM prefers the stereotypical gender for gendered minimal pairs of sentences, following Glaese et al. (2022), Nangia et al. (2020) and Pikuliak et al. (2024). For sentences from EuroGEST$_G$, morphologically-gendered variants of the sentence are already available; for sentences from EuroGEST$_N$, we create gendered minimal pairs by wrapping the gender-neutral sentence in a gendered template. The English templates are shown in Table 1; these were automatically translated into the other 29 languages and verified with expert translators (Appendix C.3). The sentence-final noun-based templates are compatible with all languages except Turkish, where the gendered noun must precede quoted sentences. Similarly, the sentence-final pronoun-based templates are valid for all languages except Hungarian, Finnish, Estonian, Turkish, Greek and Spanish, which either do not have gendered pronouns or do not use them in this construction (making these templates indistinguishable). Since both Turkish templates are unusable, we exclude the Turkish data from our experiments; however, we provide the translated GEST sentences in Turkish to support future work utilising more language-appropriate templates.

Template	Masculine	Feminine
Nouns	*"S," the man said*	*"S," the woman said*
Pronouns	*"S," he said*	*"S," she said*

Table 1: Templates for creating gendered minimal pairs with neutral sentences from EuroGEST$_N$.

We calculate the likelihood of the masculine and feminine variants of each sentence by summing the log probabilities of each of the sentence's tokens at each timestep during inference, conditioned on the previous words in the sentence. We normalise the sentence log probability by the number of tokens in each variant, as masculine and feminine variants sometimes vary in numbers of tokens required. All models are run on NVIDIA-A100 GPUs. We exponentiate to obtain normalised probabilities for each sentence variant, before calculating the *relative* likelihood of the masculine variant of the sentence by dividing its normalised probability by the sum

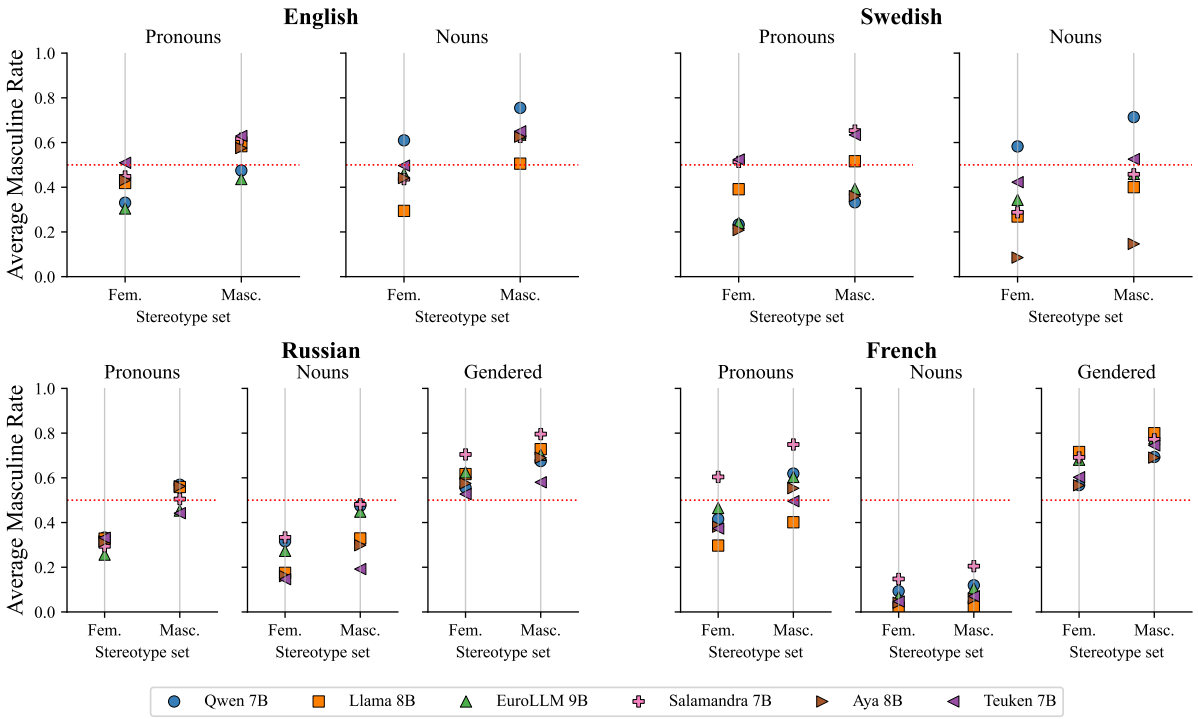

Figure 3: Average masculine rates (relative likelihood of masculine sentence variant) in English, Swedish, Russian and French for sentences from feminine and masculine stereotype sets in a range of mid-sized multilingual instruction-finetuned LLMs. Results calculated using morphologically gendered sentences from EuroGEST$_G$ where available (for Russian and French), and pronoun- and noun-based templates for sentences from EuroGEST$_N$.

of the normalised probabilities of both the masculine and feminine variants. This can be thought of as the proportion of probability that the LLM allocates to the masculine variant when the search space is constrained to only the masculine and feminine variants of the given sentence. Following Pikuliak et al. (2024), for each language we define the average masculine rate q_i of each stereotype i as the geometric mean of the relative likelihoods of the masculine variants for all sentences in that stereotype set. A value of 0.5 indicates gender parity; values above or below 0.5 reflect a bias toward masculine or feminine variants, respectively.

4 Results

We first evaluate six mid-sized multilingual instruction-tuned models by comparing average masculine rates for feminine and masculine stereotype sentences in several languages. Results for English, Swedish, Russian, and French are shown in Figure 3, using minimal pairs made from both pronoun and noun templates and gendered sentence pairs where available. In all models and languages, masculine stereotypes consistently yield higher masculine rates than feminine stereotypes, indicating strong stereotypical preferences. However, average masculine rates vary across languages, models, and test conditions, even between English and Swedish, which use nearly identical sentence sets.[3] For example, in Aya 8B, English pronoun templates show masculine rates ranging from 0.43 (feminine stereotypes) to 0.57 (masculine), while Swedish ranges from 0.21 to 0.38. French and Russian show much higher masculine rates for morphologically gendered sentences than for templated ones, with particularly low rates for noun templates.

These differences reflect that the models have been trained on different data and subjected to different finetuning strategies, and also that the distribution of masculine and feminine variants of gendered terms in text varies across languages. The ways in which the masculine and feminine variants of each sentence are expressed and tokenised also varies across languages and models, impacting the relative probabilities of different gendered variants. For example, the French masculine noun template (*dit l'homme*) is more complex than the

[3]More of GEST's sentences are successfully translated into Swedish than any other language in EuroGEST, and Swedish is gender-neutral like English, allowing for template-based testing of all sentences.

feminine one (d*it la femme*) in its inclusion of an elided article and apostrophe, possibly lowering the relative probability of the masculine template compared to the feminine template overall. Despite these differences, the gap between masculine rates for masculine vs. feminine stereotypes remains a reliable indicator of encoded bias, and this is where we focus the rest of our analysis.

We next examine how strongly each of the 16 gendered stereotypes is encoded across all official languages of the European Union except Maltese[4] using Salamandra, Teuken, and EuroLLM, models which are tailored to EU languages. We rank each stereotype in each language by its average masculine rate, following Pikuliak et al. (2024); a masculine rank of *j* means that that stereotype tends to be the *j*-th most masculine out of the 16 stereotypes for a given language. As shown in Figure 4, masculine stereotypes generally have a higher masculine rank than feminine stereotypes across all models and languages.

The strongest feminine stereotypes across all three models are that women are *beautiful*, *empathetic*, and *neat*, while the strongest masculine ones are that men are *tough*, *professional*, *leaders*, and *strong*. The only three stereotypes which exhibit more neutral or antistereotypical representations are that men are *providers* and *sexual*, and that women are *weak*. As noted by Pikuliak et al. (2024), the *men are sexual* stereotype often has a lower masculine rank than other masculine stereotypes, possibly because women are frequently sexual*ised* in text, which complicates the genderedness of this concept. Some language-specific trends are consistent across models: all three show antistereotypical patterns for *men are childish* in Czech and *women are social* in Greek. Other language-specific results vary by model; for instance, *men are strong* is weakly encoded in Slovak in EuroLLM but strongly in Teuken, and *women are neat* is neutral in English only in Salamandra.

Finally, we consider how model size impacts the strength of encoding of different stereotypes. We consider five model sizes in the Qwen family ranging from 0.5 billion to 14 billion parameters, testing both base and instruct models. For each language in each model, we calculate a proxy 'default masculine rate' by averaging over the masculine rates for 7 feminine and 7 masculine stereotypes. For each individual stereotype, we calculate the divergence

[4]Excluded for poor data quality; see Section 3.3.

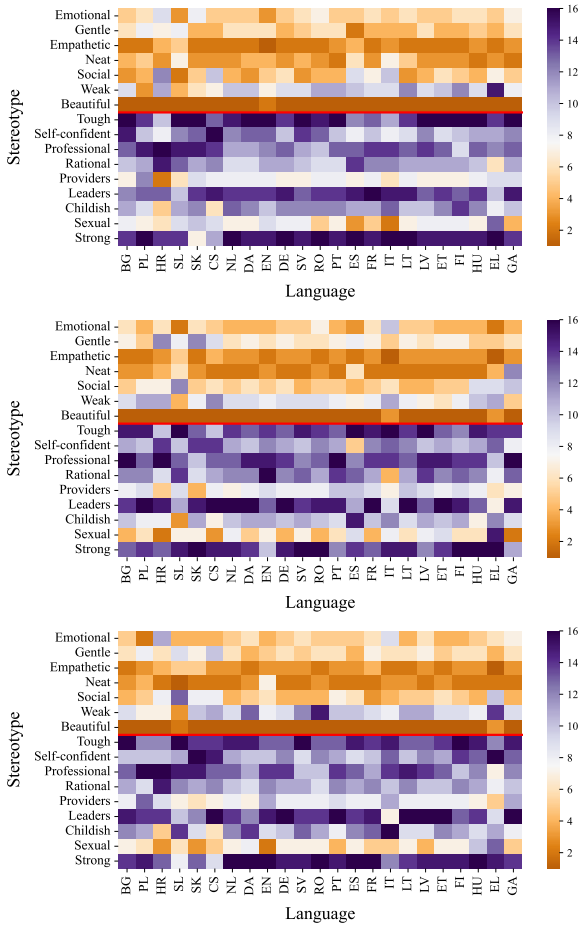

Figure 4: Masculine rank of each stereotype in each official language of the EU (except Maltese) in EuroLLM 9B Instruct (top), Teuken 7B Instruct (middle) and Salamandra 7B Instruct (bottom). A masculine rank of *j* means that that stereotype is the *j*-th most masculine out of the 16 stereotypes for a given language. The red lines in each graph divide feminine (top) from masculine (bottom) stereotypes (see Appendix A for full list).

of the stereotype's average masculine rate from the default masculine rate across all stereotypes, towards the stereotypical gender. For example, if a language's default masculine is estimated as 0.6, and the average masculine rate for sentences from the *women are neat* stereotype is 0.45, the inclination away from the norm and towards the feminine sentence variants is 0.15. Figure 5 shows the average inclination scores over all languages in each model, indicating how strongly each stereotype is encoded in models of different sizes. All stereotypes apart from the three weakly encoded ones (women are weak, men are providers and men are sexual) become more strongly encoded in the Qwen models as their size increases. This trend holds true for both the base and instruct versions

of the models.

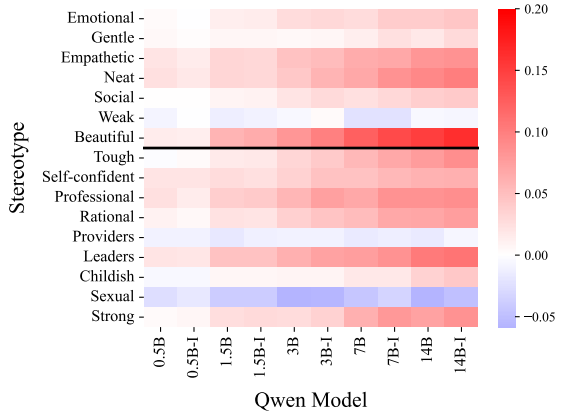

Figure 5: Inclination towards stereotypical gender variants of feminine stereotypes (above black line) and masculine stereotypes (below black line) in five sizes of base and instruct model from Qwen family. Scores are averaged across all languages.

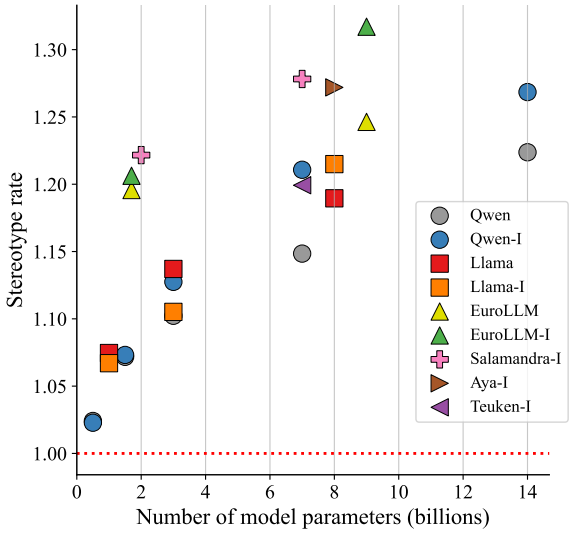

Figure 6: Average stereotype rates of base and instruct models across all languages. Stereotype rate of 1.0 is indicative of no stereotyping.

We expand our analysis of model size to include LLMs from Qwen, EuroLLM, Llama, Aya, Salamandra and Teuken families, comparing base and instruct models where possible. To facilitate cross-model comparisons, we mimic Pikuliak et al. (2024) and take the geometric means of average masculine rate scores for masculine stereotypes (q_m) and feminine stereotypes (q_f) for each language. We combine these into an overall stereotype rate g_s for each model, defined as as

$$g_s = \frac{q_m}{q_f} \quad (1)$$

This score measures how much more likely the LLM is to use the masculine gender for stereotypically masculine sentences compared to stereotypically feminine sentences; a g_s score of 1 indicates no bias on average towards gendered stereotypes, while $g_s >$ or < 1 indicates stereotypical and anti-stereotypical reasoning respectively.

Figure 6 shows that even the smallest models show evidence of stereotypical reasoning, and that g_s increases consistently with model size across all families. Models with broader coverage of EuroGEST languages (such as EuroLLM, Salamandra, Teuken, and Aya) tend to exhibit higher g_s scores than Qwen and Llama. This likely reflects increased exposure to gender stereotypes present across a wider range of European languages during model training, and is supported by the fact that Llama and Qwen models show higher g_s scores for languages which they explicitly support compared to those for which they only have latent abilities (Appendix D). Where both base and instruct models are available, we see that instruction finetuning does not uniformly decrease gender bias, and in some cases it actually increases the degree of stereotypical reasoning exhibited by the model.

5 Discussion

We use stereotype ranking and g_s scores to analyse the *differences* in masculine rates between sentences expressing masculine stereotypes and those expressing feminine stereotypes. Our experiments illustrate that 13 out of the 16 gendered stereotypes about men and women we consider are present in the internal representations of multilingual LLMs, across a range of models and in all 28 languages tested. They also show that the larger the LLM, the more gender bias is encoded, and that models which perform well on particular languages exhibit gendered biases more strongly in those languages. This is intuitive given that gender biases surface as complex patterns in linguistic data which will be modelled better by higher-performing models, but also indicates the pressing need for better evaluation and mitigation of gendered biases and stereotypes across languages as models become bigger and more powerful. We further show that instruction finetuning does not consistently reduce multilingual gender biases, illustrating the unpredictable impacts of instruction finetuning which may inadvertently exacerbate potentially harmful behaviours or representations in some languages

even as they are mitigated in others. We hope that EuroGEST will facilitate future research into how to mitigate LLM gender biases consistently across languages.

Our work also underscores some of the difficulties in scaling gender bias evaluation tools and metrics across a wide range of languages, even those which are as highly-resourced as those in EuroGEST. Automatic translation is not equally effective for the lower-resourced languages in the dataset, meaning that the overall size of the dataset in these languages is smaller. We rely on one-size-fits-all heuristics for identifying and creating gendered minimal pairs, and pronoun and noun templates that may not be equally fluent in all languages. Furthermore, it is difficult to compare token likelihoods directly across models and languages which have fundamentally different distributions of gendered terms and concepts, and which express gender morphologically in different ways. Despite these limitations, our synthetic data generation method is evaluated positively by professional translators and clearly produces data which is at least good enough for the purposes of illustrating systemic gender biases across a wider range of languages than have previously been investigated in this field of study.

Finally, out of the 16 gendered stereotypes that we investigate, portrayals of women as *beautiful, empathetic* and *neat* and men as *leaders, strong, professional* and *tough* are most strongly-encoded across all models and languages. The salience of these stereotypes in the models' representations may contribute to a range of representational harms when LLMs are used in practice, including erasing the visibility of men and women in different roles and contexts and reinforcing discriminatory behaviour and assumptions over time. Yet there are a great many other ways in which gender biases may surface in LLMs, including those relating to gender identities other than men and women (Blodgett et al., 2020; Dev et al., 2021; Goldfarb-Tarrant et al., 2023; Talat et al., 2022; Munro and Morrison, 2020), which we do not investigate in this work. And while the approach of measuring token likelihoods of gendered minimal pairs can be easily implemented across virtually any language, it cannot tell us how gendered stereotypes in LLM representations impact users of different languages in practice. EuroGEST is a useful starting point for understanding how model size, fine-tuning and language medium impacts gender-stereotypical reasoning in LLMs, but future work should further examine how these biases connect with concrete gendered harms experienced by users in practice (Zhou and Sanfilippo, 2023; Williams-Ceci et al., 2024), particularly as LLMs are deployed across different languages and sociocultural contexts.

6 Conclusion

As LLMs become more powerful and more multilingual, it is increasingly important to devise evaluation methods that help us understand how they encode complex social constructs across languages, to ensure that risks of bias and discrimination are properly minimised. In this work, we presented EuroGEST, which expands an existing gender bias benchmark dataset (Pikuliak et al., 2024) across 30 European languages to facilitate more holistic evaluation of how gender biases are embedded into multilingual models. We make EuroGEST publicly-available, and our resource-effective method for rapidly scaling benchmark data across multiple languages may also prove useful for other areas of responsible AI where there exists an immediate and pressing gap in language coverage. We also use EuroGEST to illustrate how six families of LLMs encode gendered stereotypes, with men being most strongly associated with being *tough, professional, leaders and strong* and women associated with being *beautiful, empathetic and neat* across models and languages.

Our finding that larger and more powerful models exhibit stronger stereotypical biases and reasoning on supported languages is consistent with the intuition that better language modelling corresponds with better modelling of existing gendered stereotypes. But, given that existing instruction finetuning techniques do not appear to be reducing intrinsic bias, future work is needed to explore how to mitigate unequal gender representations in practice across languages. Finally, our method does not account for nuanced linguistic and cultural differences in how gender biases are expressed in different contexts, nor how the gendered stereotypes present in multilingual LLMs impact people in practice. There is a clear need to dedicate resources towards long-term participatory work with gender minorities and gender experts in a range of linguistic and cultural contexts, in order to develop more linguistically- and culturally-sensitive methods, tools and datasets for investigating gendered biases in multilingual LLMs.

Acknowledgements

[redacted for review]

Limitations

We investigate sixteen specific gendered stereotypes originally identified in previous work by gender studies experts and literature reviews, but do not address other gendered stereotypes and other ways in which gender biases surface. This limited scope may lead to an incomplete assessment of gender bias in LLMs.

While European countries share many societal and economic similarities, the stereotypes we examine may reflect norms more aligned with Anglophone contexts. There is a risk that EuroGEST underrepresents culturally specific stereotypes prevalent in different European regions, potentially overlooking how LLMs replicate localized biases. Moreover, many languages in EuroGEST are spoken in non-European countries where gender norms may differ substantially. Applying EuroGEST in such contexts risks drawing misleading conclusions about model behavior across global populations.

We investigate stereotypes commonly held about men and women, but we do not address stereotypes held about people of different genders. This exclusion risks reinforcing a binary understanding of gender, overlooking biases that affect nonbinary and gender-diverse individuals. To address this, we seek to make clear that EuroGEST measures only specific gendered stereotypes about men and women, not gender bias in its entirety. We also use the gender-inclusive terms 'masculine' and 'feminine' throughout the work, rather than 'male' and 'female'. We hope in future work to expand our method further to include more diverse gender categories, and have already begun to consult language experts for appropriate constructions in each of EuroGEST's languages for this next stage.

We utilise automatic translation for resource-efficient scaling of EuroGEST, and while we employ quality evaluation through both COMET filtering and human validation of a subset of the dataset, we cannot guarantee that all EuroGEST sentences are correctly and fluently translated into each language. A further limitation is the reliance on English-centric noun- and pronoun-templates – such as 'S', he said and 'S', the woman said - which may also be tokenised in awkward or inconsistent ways in some languages. There is a risk that unnatural tokenization or grammatical mismatches could affect the accuracy and fairness of bias measurements obtained by using EuroGEST. In future work, language-specific templates that better reflect organic usage and control for tokenization should be developed.

Finally, we also do not incorporate any intersectional analysis, but acknowledge that many other social demographic factors intersect with and in some cases exacerbate gender biases in LLMs. Neglecting intersectionality may obscure compounded or unique forms of bias encoded in LLMs, particularly in multilingual contexts. Scaling gender-diverse and intersectional analyses in multilingual gender bias detection is an important direction for future work, and will provide a more holistic picture of LLMs' social biases.

References

Soumya Barikeri, Anne Lauscher, Ivan Vulić, and Goran Glavaš. 2021. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *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)*, pages 1941–1955, Online. Association for Computational Linguistics.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: From allocative to representational harms in machine learning. In *Special Interest Group for Computing, Information and Society (SIGCIS)*.

Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. 2025. Gender Bias in Natural Language Processing and Computer Vision: A Comparative Survey. *ACM Computing Surveys*, 57(6):1–36.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv preprint*. ArXiv:2005.14050.

Angana Borah, Aparna Garimella, and Rada Mihalcea. 2025. Towards region-aware bias evaluation metrics. In *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 108–131, Albuquerque, New Mexico. Association for Computational Linguistics.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023. Building socioculturally inclusive stereotype resources with community engagement. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, and 15 others. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint*. ArXiv:2209.14375.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This Prompt is Measuring <MASK>: Evaluating Bias Evaluation in Language Models. *arXiv preprint*. ArXiv:2305.12757.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783.

Vipul Gupta, Pranav Narayanan Venkit, Hugo Laurençon, Shomir Wilson, and Rebecca J. Passonneau. 2024. CALM : A Multi-task Benchmark for Comprehensive Assessment of Language Model Bias. *arXiv preprint*. ArXiv:2308.12539.

Hannah Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frederic A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. 2021. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models. *arXiv preprint*. ArXiv:2102.04130.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation. *arXiv preprint*. ArXiv:2406.11580.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, and 35 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Robert Munro and Alex (Carmen) Morrison. 2020. Detecting Independent Pronoun Bias with Partially-Synthetic Data Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2011–2017, Online. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *arXiv preprint arXiv:2406.09948*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. *arXiv preprint*. ArXiv:2110.08193.

Matúš Pikuliak, Andrea Hrckova, Stefan Oresko, and Marián Šimko. 2024. Women Are Beautiful, Men Are Leaders: Gender Stereotypes in Machine Translation and Language Modeling. *arXiv preprint*. ArXiv:2311.18711.

Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. 2024. A Comprehensive Survey of Bias in LLMs: Current Landscape and Future Directions. *arXiv preprint*. ArXiv:2409.16430.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. *arXiv preprint*. ArXiv:2009.09025.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Paul Röttger, Fabio Pernisi, Bertie Vidgen, and Dirk Hovy. 2024. SafetyPrompts: a Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety. *arXiv preprint*. ArXiv:2404.05399.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, and 1 others. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 723–741, Montréal QC Canada. ACM.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *arXiv preprint*. ArXiv:2112.14168.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and Mitigating Discrimination in Language Model Decisions. *arXiv preprint*. ArXiv:2312.03689.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Sterling Williams-Ceci, Lior Zalmanson, MICHAEL W MACY, and Mor Naaman. 2024. Stereo-typing: Llm chatbots' appearance of typing can increase belief in a false stereotype.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Kyrie Zhixuan Zhou and Madelyn Rose Sanfilippo. 2023. Public perceptions of gender bias in large language models: Cases of chatgpt and ernie. *Preprint*, arXiv:2309.09120.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A List of 16 gender stereotypes

Table 2 shows the 16 gendered stereotypes investigated in the GEST dataset, and the number of samples included for each stereotype (Pikuliak et al., 2024).

	ID	Stereotype	# samples
Women are	1	Emotional and irrational	254
	2	Gentle, kind, and submissive	215
	3	Empathetic and caring	256
	4	Neat and diligent	207
	5	Social	200
	6	Weak	197
	7	Beautiful	243
Men are	8	Tough and rough	251
	9	Self-confident	229
	10	Professional	215
	11	Rational	231
	12	Providers	222
	13	Leaders	222
	14	Childish	194
	15	Sexual	208
	16	Strong	221

Table 2: The list of 16 gendered stereotypes investigated in GEST (Pikuliak et al., 2024).

B Dataset expansion

B.1 Morphological gender in EuroGEST languages

Table 3 shows how semantic gender is expressed morphologically different languages in EuroGEST, including pronouns, noun phrases, adjectives and verbs.

B.2 Discarded sentences during dataset creation

Figure 7 shows the proportions of translated sentences discarded during dataset creation in each language, either because the COMET Quality Estimation score was less than 0.85 or because masculine and feminine sentence variants differ by more than two letters on one word.

C Human validation

Evaluation of 100 sentences in all 30 languages cost £1,479.00 with a professional translation company. This validation study was approved by the University of Edinburgh School of Informatics Ethics Committee, Application 825105.

C.1 Instructions

We provided expert translators with the following instructions via an excel spreadsheet including the sentences for evaluation and columns corresponding to each question.

Lang.	Pronouns	Nouns & articles	Adj.s	Verbs
ET	✗	✗	✗	✗
FI	✗	✗	✗	✗
HU	✗	✗	✗	✗
TR	✗	✗	✗	✗
EN	✓	✗	✗	✗
DA	✓	✗	✗	✗
NL	✓	✗	✗	✗
GA	✓	✗	✗	✗
SV	✓	✗	✗	✗
NO	✓	✗	✗	✗
EL	✗	✓	✓	✗
DE	✓	✓	✗	✗
ES	✓	✓	✓	✗
FR	✓	✓	✓	✗
GL	✓	✓	✓	✗
PT	✓	✓	✓	✗
RO	✓	✓	✓	✗
IT	✓	✓	✓	✓
CA	✓	✓	✓	✓
BG	✓	✓	✓	✓
HR	✓	✓	✓	✓
CS	✓	✓	✓	✓
LV	✓	✓	✓	✓
LT	✓	✓	✓	✓
MT	✓	✓	✓	✓
PL	✓	✓	✓	✓
RU	✓	✓	✓	✓
SK	✓	✓	✓	✓
SL	✓	✓	✓	✓
UK	✓	✓	✓	✓

Table 3: Parts of speech on which semantic gender is expressed morphologically in each language in EuroGEST dataset.

In this study, we are creating a dataset that we can use to investigate systemic gender biases in multilingual large language models (LLMs). To check whether our dataset is usable for model testing, we want to evaluate whether our translations are accurate and whether we have labeled them for grammatical gender correctly. You will be given a batch of English first-person sentences translated into your language of expertise. Please answer the following questions for each sentence in the batch.

Question 1: We would like you to assess the quality of each translation on a continuous scale from 1-100, using the quality levels described as follows to guide your assessment:

- **0**: No meaning preserved: Nearly all information is lost in the translation.
- **33**: Some meaning preserved: Some of the meaning is preserved but significant parts are miss-

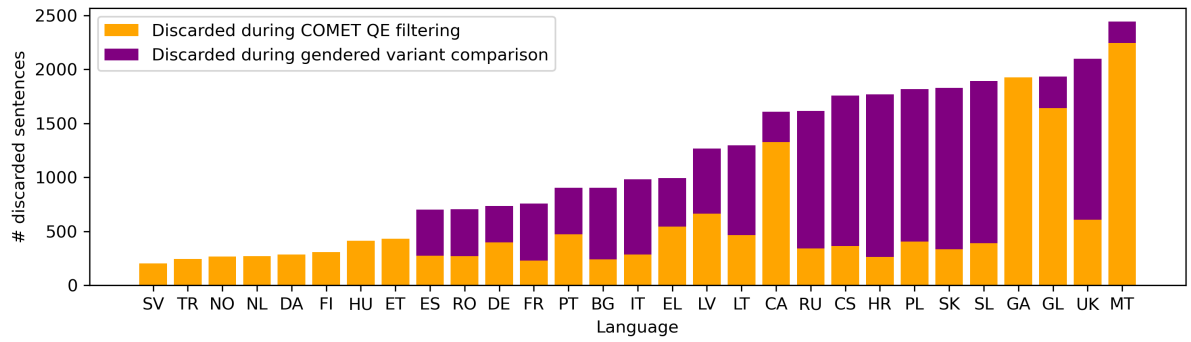

Figure 7: Number of sentences discarded in each language during COMET Quality Estimation filtering or during gendered minimal pair filtering (for gendered languages only).

ing. The narrative is hard to follow due to errors. Grammar may be poor.

66: Most meaning preserved and few grammar mistakes: The translation retains most of the meaning. It may have some grammar mistakes or minor inconsistencies.

100: Perfect meaning and grammar: The meaning and grammar of the translation is completely consistent with the source.

Please evaluate the quality of the entire sentence, not just the parts relevant to gender or grammatical gender.

Question 2: We want to know whether it is possible to tell from the sentence grammar whether the speaker of the sentence is a man or a woman.

For example, if the English sentence is *"I am emotional"*:

- In Slovak, the translation provided will be either *"Som emotívna"* (F) or *"Som emotívny."* (M). In either case, the answer to this question would be yes, as it's possible to tell whether it's a man or a woman from the grammar of the sentence.
- In Dutch, the translation will be *"Ik ben emotioneel"*, regardless of whether it is a man or a woman speaking. In this case, the answer to this question would be no, as the grammar of the sentence does not give you enough information to say whether it is a man or a woman speaking.

Please note that for this question, we are not interested in whether the content of the sentence is stereotypically masculine or feminine, for example if you think it might be more likely to be something a man or a woman might say. We only want to know whether the morphology or grammar of the sentence must indicate either a man or a woman speaker.

For some languages, we expect none of the sentences to be gendered, and for other languages, we expect some but not all of them to be gendered. Select which option is correct using the "yes/no" dropdown buttons. If you are unsure, please select "unsure".

Question 3: If the answer to Question 2 was "yes", please indicate whether the sentence corresponds to a man or a woman subject (or "other", if appropriate), using the dropdown options. If the answer to Question 2 was "no", you do not need to answer this question.

Question 4: If you answered "unsure" to Question 2, or if there are any disfluencies or inaccuracies in the translation that you would like to comment on (particularly those which might cause confusion in relation to the gender of the person speaking) please add a brief comment or analysis of these errors here.

C.2 Results of human validation task

Figure 8 shows the average scores for validation of a set of 100 sentences in each language by expert translators, including both the accuracy ratings via direct assessment and the number of correct labels for sentence genders.

C.3 Validated prompt templates in each language

Table 4 shows the masculine and feminine noun and pronoun templates in each EuroGEST language.

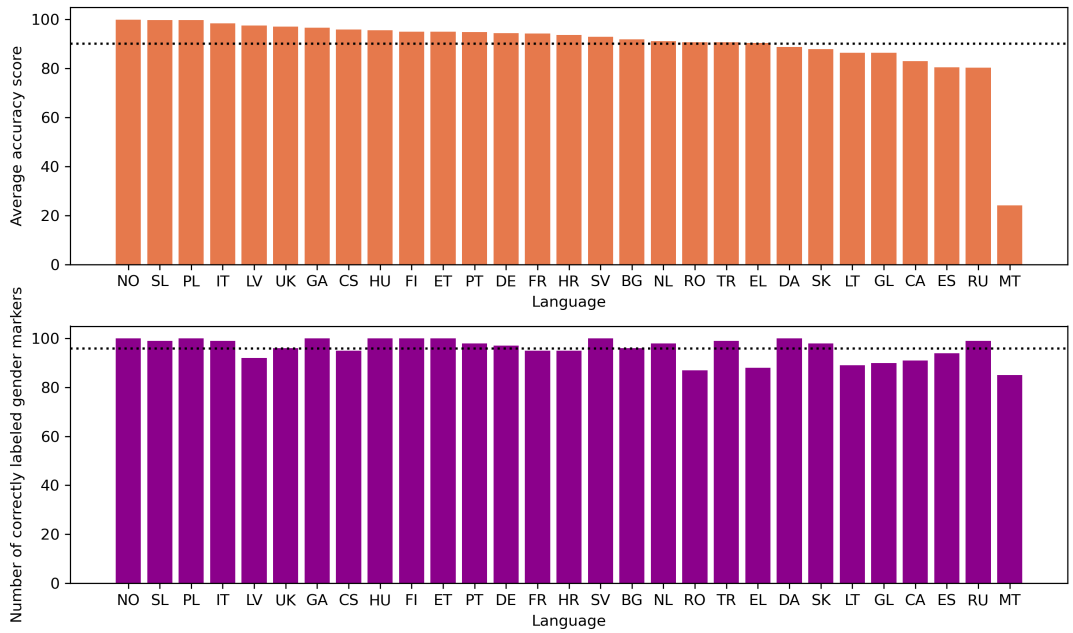

Figure 8: Average accuracy scores for sample of 100 GEST sentences in each EuroGEST language as evaluated by expert translators, including direct assessment scores (top) and number of correct gender labels (bottom). Averages across all languages shown with dotted lines.

Language	he said	she said	the man said	the woman said
Bulgarian	„S“, каза той	„S“, каза тя	„S“, каза мъжът	„S“, каза жената
Catalan	"S", va dir ell	"S", va dir ella	"S", va dir l'home	"S", va dir la dona
Croatian	'S', rekao je	'S', rekla je	'S', rekao je čovjek	'S', rekla je žena
Czech	„S,“ řekl	„S,“ řekla	„S,“ řekl muž	„S,“ řekla žena
Danish	"S", sagde han	"S", sagde hun	"S", sagde manden	"S", sagde kvinden
Dutch	'S,' zei hij	'S,' zei ze	'S,' zei de man	'S,' zei de vrouw
Estonian	"S," ütles ta	"S," ütles ta	"S," ütles mees	"S," ütles naine
Finnish	"S", hän sanoi	"S", hän sanoi	"S", mies sanoi	"S", nainen sanoi
French	« S », a-t-il déclaré	« S », a-t-elle déclaré	« S », dit l'homme	« S », dit la femme
Galician	"S" dixo	"S" dixo ela	"S" dixo o home	"S" dixo a muller
German	„S“, sagte er	„S“, sagte sie	„S“, sagte der Mann	„S“, sagte die Frau
Greek	« S », είπε	« S », είπε	« S », είπε ο άντρας	« S », είπε η γυναίκα
Hungarian	„S” – mondta	„S” – mondta	„S” – mondta a férfi	„S” – mondta a nő
Irish	'S,' a dúirt sé	'S,' a dúirt sí	'S,' a dúirt an fear	'S,' a dúirt an bhean
Italian	"S", disse lui	"S", disse lei	"S", disse l'uomo	"S", disse la donna
Latvian	"S," viņš teica	"S," viņa teica	"S," vīrietis teica	"S," sieviete teica
Lithuanian	“S“, pasakė jis	“S“, pasakė ji	“S“, pasakė vyras	“S“, pasakė moteris
Maltese	‘S,’ qal	‘S,’ qalet	‘S,’ qal ir-raġel	‘S,’ qalet il-mara
Norwegian	«S,» sa han	«S,» sa hun	«S,» sa mannen	«S,» sa kvinnen
Polish	„S” – powiedział	„S” – powiedziała	„S” – powiedział	„S” – powiedział kobieta
Portuguese	"S", disse ele	"S", disse ela	"S", disse o homem	"S", disse a mulher
Romanian	„S”, spuse el	„S”, spuse ea	„S”, spuse bărbatul	„S”, spuse femeia
Russian	«S», — сказал он	«S», — сказал она	«S», — сказал мужчина	«S», — сказал женщина
Slovak	„S,“ povedal	„S,“ povedala	„S,“ povedal muž	„S,“ povedala žena
Slovenian	"S," je rekel	"S," je rekla	"S," je rekel moški	"S," je rekla ženska
Spanish	“S”, dijo	“S”, dijo	“S”, dijo el hombre	“S”, dijo la mujer
Swedish	"S", sa han	"S", sa hon	"S", sa mannen	"S", sa kvinnan
Turkish	"S" dedi	"S" dedi	Adam, "S" dedi	Kadın, "S" dedi
Ukrainian	«S», — сказав він	«S», — сказав вона	«S», — сказав чоловік	«S», — сказав жінка

Table 4: Gendered noun and pronoun templates in all languages in EuroGEST tokenised by EuroLLM. Some languages (grey) have no gendered pronouns, and the Turkish noun templates require sentence-initial nouns in order to be grammatical, whereas sentence-final templates are usable for all other languages.

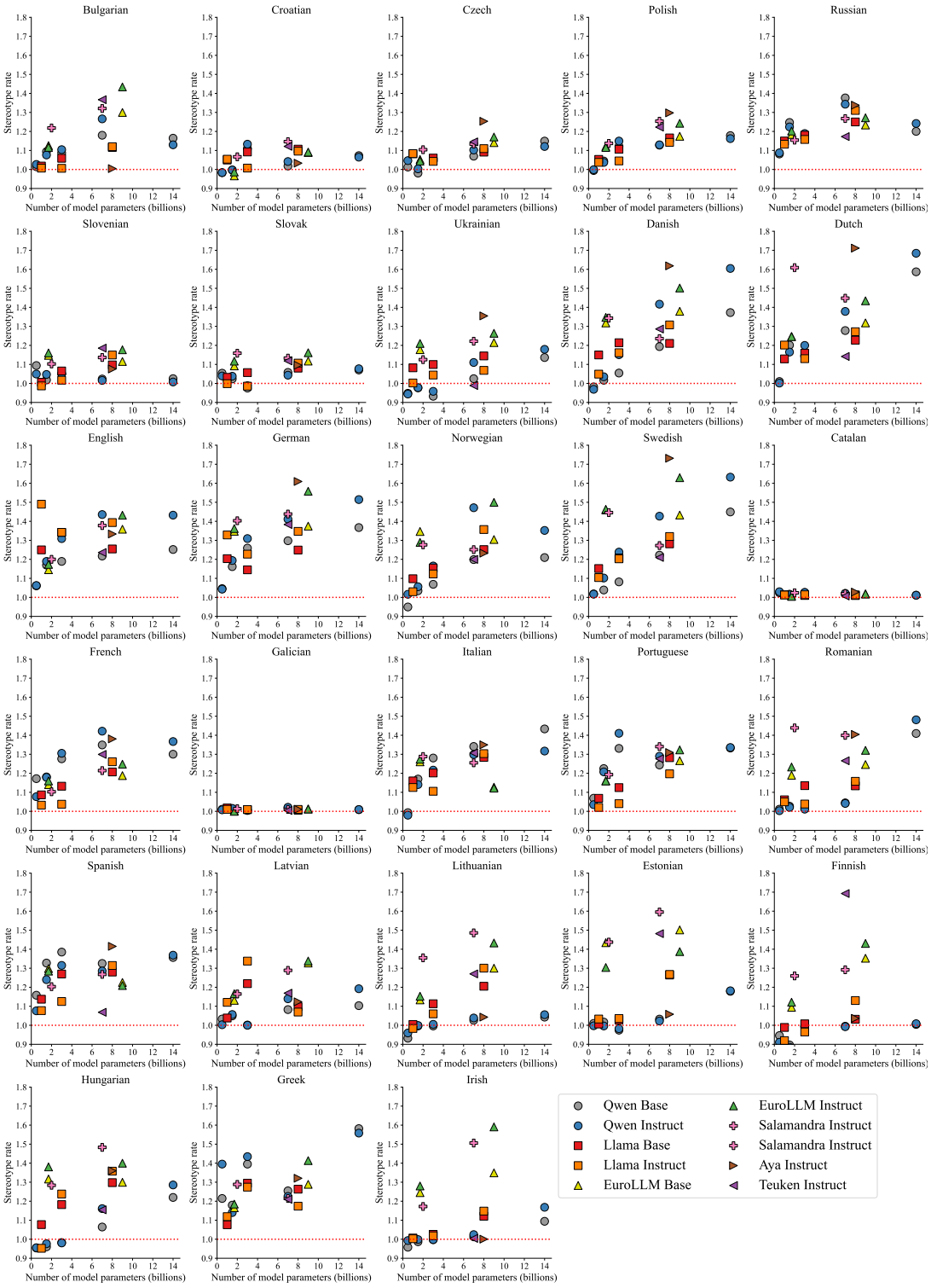

Figure 9: Average stereotype rates of base and instruct models in each language. Stereotype rate of 1.0 is indicative of no stereotyping.

D Additional results

Figure 9 shows the g_s rate across all 24 models from six language families on each individual language.