EuroGEST: Investigating gender stereotypes in multilingual language models

Jacqueline Rowe[1], Mateusz Klimaszewski[2], Liane Guillou[3], Shannon Vallor[1], Alexandra Birch[1,3]

[1]University of Edinburgh, [2]Warsaw University of Technology, [3]Aveni

Correspondence: jacqueline.rowe@ed.ac.uk

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 finetuned models continue to exhibit 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 help identify and quantify systemic biases in LLMs, but their utility depends on clearly articulating the motivations, values and norms embedded in their design (Blodgett et al., 2020; Goldfarb-Tarrant et al., 2023).

Most existing bias benchmarks serve a limited number of languages (Blodgett et al., 2020; Röttger et al., 2024), and widely-used multilingual LLMs are not consistently evaluated for bias across all supported languages (see, for example, Grattafiori et al. (2024); Martins et al. (2024); NLLB Team et al. (2022); Üstün et al. (2024)). Consequently, there is little understanding of how social biases in LLMs vary across languages, and these inconsistencies in bias evaluation methods may result in discriminatory outcomes when LLMs are deployed in multilingual contexts. The current lack of multilingual bias evaluation tools also makes it difficult to assess the cross-lingual effectiveness of (largely English-centric) bias mitigation techniques for LLMs across languages.

In this work, we explore gendered stereotyping by multilingual generative LLMs. Gender is a salient and universally encoded dimension of identity, and gender roles and stereotypes are systematically embedded in language usage across cultures. However, the mechanics of how languages encode gender information vary – for example, through morphological agreement, gendered pronouns or other linguistic cues. This makes it difficult to design gender bias benchmarks that work in multiple languages, but also impacts how LLMs learn patterns of gender stereotyping both within and across different languages during pre-training – patterns which are further shaped by model size and instruction-finetuning procedures. Together, these factors motivate our three research questions:

1. *How can we leverage machine translation technologies to make more multilingual gender bias benchmarks?*

2. *Do multilingual LLMs exhibit consistent gender stereotyping patterns across languages?*

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32074-32096
November 4-9, 2025

3. *How does model size or instruction finetuning affect the degree of gender stereotyping exhibited by different families of LLMs?*

To explore these questions, we introduce EuroGEST,[1] a new gender bias benchmark dataset that adapts and extends an existing open-source multilingual gender bias benchmark dataset (Pikuliak et al., 2024) to cover 29 European languages from five major language families.[2] We focus on European languages because they are relatively highly-resourced, facilitating automatic scaling of benchmark data via machine translation. Cultural and socio-economic parallels across Europe also make gender stereotypes within European countries more comparable than between European and non-European contexts. Our main contributions are as follows:

- **Benchmark creation (RQ1)**: We develop an automated pipeline for generating gendered minimal pairs of sentences in different languages, using it to create a novel dataset of 71,000 sentences linked to 16 gendered stereotypes across 30 European languages;
- **Bias evaluation (RQ2, RQ3)** We use the novel dataset to evaluate 24 multilingual LLMs for gendered stereotyping across all 30 languages, demonstrating that stereotyping increases with larger model sizes, across both base and instruction-finetuned models.

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, especially across different languages and cultures. Here we summarise existing techniques and highlight gaps with regard to multilingual gender bias detection.

Extrinsic Bias Metrics 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 to prompts including proper names associated with different demographic groups to investigate whether the LLM exhibits bias in performance on the task based on the name identity. Tamkin et al. (2023) focus on decisionmaking tasks, creating prompt templates for investigating bias in realistic scenarios spanning finance, business, law and education. 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 genderless into gendered languages follow stereotypical biases (Savoldi et al., 2021; Stanovsky et al., 2019; Bentivogli et al., 2020; Pikuliak et al., 2024; Mastromichalakis et al., 2025). While these studies have a helpful focus on how gendered harms might arise as LLMs are utilised in practice, it can be difficult to scale such approaches to novel languages, and the translation directions that can be evaluated in this fashion are limited.

Intrinsic Bias Metrics 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) co-reference 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 com-

[1]Available at https://github.com/JacquelineRowe/EuroGEST 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.

ments. These methods do not predict whether a model will behave in a discriminatory fashion in a specific use case, but can reveal strong underlying biases that may require closer examination of their impact on performance in certain contexts.

One key advantage of intrinsic bias metrics relevant to the current work is that they can 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. Dataset samples can be curated by local and native speakers of each language context to produce multilingual benchmark data which is well-adapted to linguistic and cultural differences in how bias is expressed (Mitchell et al., 2025; Borah et al., 2025; Dev et al., 2023; Myung et al., 2024). However, this method of curation is highly-resource intensive, and there is an interim need for more rapidly scaleable methods to expand bias benchmarks across a greater range of languages to help identify and address potential representative and allocational harms.

3 Dataset Expansion

Our first research question asks how we can use machine translation technologies to build more multilingual gender bias benchmarks. Of the 30 European languages of focus, 20 are *gendered* (they express gender on adjectives, nouns or verbs), while 10 are *genderless* (expressing morphological gender only on pronouns (6 languages) or not at all (4 languages) (see Section B). The dataset we select and the methods we use to expand it across languages must account for this variability.

3.1 Dataset selection

Our work builds on the GEST (GEnder-STereotypes) dataset created by Pikuliak et al. (2024). GEST consists of 3,565 manually-generated sentences associated with 16 common gendered stereotypes about men and women (listed in Section A). Each sentence is gender-neutral in English and gendered when translated into a Slavic language; for example, *"I am emotional"* is *"Som emotívny"* (masculine) or *"Som emotívna"* (feminine), in Slovak. The authors use these sentences to test for gender bias in translation from English into Slavic languages, and in text generation for masked and generative language models. To evaluate generative LLMs, they calculate the probability of the masculine token and the feminine token at the point where the sentences differ, and examine whether the model prefers grammatically feminine versions of sentences associated with feminine stereotypes and vice versa (see Section 4). For the genderless English sentences, they apply the same method by wrapping each sentence in a gendered template (see Table 1) and comparing the likelihoods on the gendered token in the template; for example, *'"I am emotional," **he/she** said'*.

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

Table 1: Templates for creating a gendered minimal pair from a neutral sentence *S* in Pikuliak et al. (2024).

We choose to expand GEST to additional languages because of its size and because its design and construction were informed by consultations with gender experts. Furthermore, to create GEST the authors introduced heuristics for identifying gendered minimal pairs of sentences in Slavic languages, which generalise well to other gendered European languages. They also showed how GEST sentences can be used to measure gender bias in both richly-gendered and genderless languages, necessary for the languages included in this work.

3.2 Dataset translation

We translate the 3,565 English sentences from GEST into 29 different languages. We use the Google Translate API because of its strong performance on translation of low-resource languages (Zhu et al., 2024). For the nine languages that – like English – lack morphological gender in first-person sentences, we simply translate the original GEST sentence to obtain a gender-neutral sentence in the target language. We use COMET-QE (Rei et al., 2020) to evaluate translation quality, because it supports all of our target languages except Maltese, and is one of the best reference-free translation quality metrics, particularly over this set of languages (Rei et al., 2022). We add any translated sentences with COMET-QE score of at least 0.85[3]

[3]We select the highest possible QE threshold that would retain at least 1,000 sentences per language at the end of

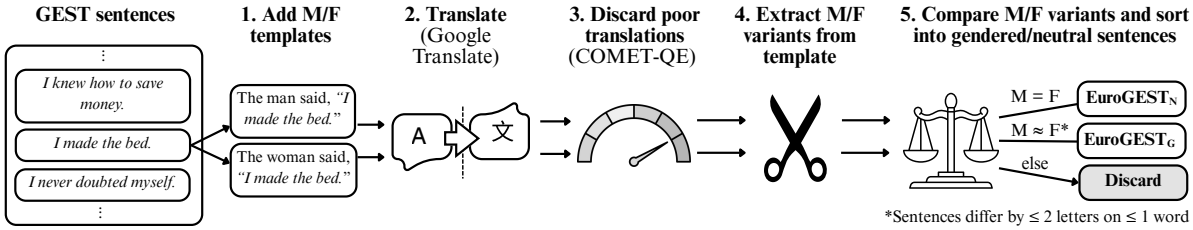

Figure 1: System for translating English GEST sentences into gendered target languages and sorting translated sentences into EuroGEST gendered (EuroGEST$_G$) and EuroGEST neutral (EuroGEST$_N$).

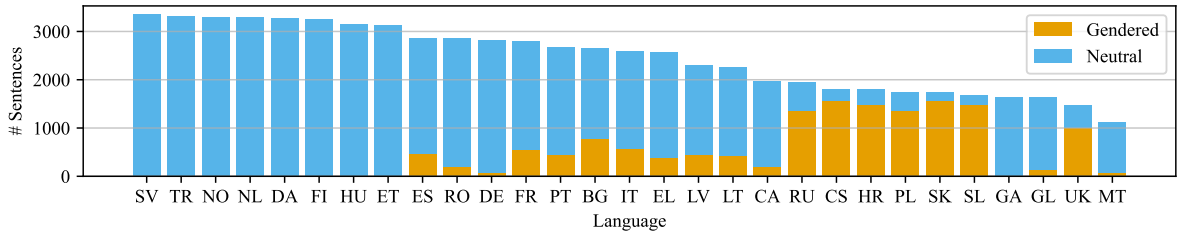

Figure 2: Number of sentences in EuroGEST-gendered and EuroGEST-neutral datasets by language.

to our EuroGEST-neutral dataset (EuroGEST$_N$), discarding the rest.

The 20 gendered languages require a more complex translation pipeline (see Figure 1) because they express gender morphologically on some but not all of the GEST sentences. For example, Italian is a gendered language, but *'I started my own company when I was 18'* is genderless (*'Ho fondato la mia azienda quando avevo 18 anni'*) while *'I gave up easily without a fight'* is gendered (*'Mi sono arres**o/a** facilmente, senza combattere'*). We account for this variation by wrapping each English GEST sentence *S* in sentence-initial masculine and feminine templates (*'The man/woman said "S"'*) and translating *both* variants into all 20 gender-sensitive languages. We apply COMET-QE filtering as before, and then extract the masculine and feminine translations of each original GEST sentence from the templated translations. If the masculine and feminine translations of the GEST sentence are identical, we assume that this GEST sentence is genderless in that language, and add it to EuroGEST$_N$. Following Pikuliak et al. (2024), if the translations differ by up to two letters on one word, we assume that they are a gendered minimal pair and add both variants to the EuroGEST-gendered dataset (EuroGEST$_G$). If the two translations differ by more than this, we discard them.[4]

With this method, we obtain 14,538 pairs of gendered sentences across the 20 gendered languages in EuroGEST$_G$, and 56,497 genderless sentences across all 30 languages (including English) in EuroGEST$_N$. During COMET-QE filtering, the most sentences are discarded from Maltese (which is not supported in COMET-QE). High numbers of sentences are also discarded from low-resource languages like Catalan, Irish and Galician at this stage, indicative of poorer translation and evaluation performance by Google Translate and COMET-QE on these languages. During identification of gendered minimal pairs of sentences in gendered languages, the most sentences are discarded from Slavic languages, likely because many of these languages express gender on more than one word in some sentences and therefore fail the strict heuristic for filtering out gender minimal pairs. Of the original GEST sentences, between 1,120 and 3,360 sentences are retained for each target language (Figure 2), averaging 155 sentences per stereotype per language (Figure 7).[5]

3.3 Validation

We validate our automatically translated and labelled data with expert translators to ensure its reliability for measuring gender stereotypes in LLMs. We select 100 sentences from each language's

dataset creation.

[4]While some languages may express gender on more than two letters in one word in a single sentence, we replicate Pikuliak et al. (2024)'s heuristic because we prefer to over-discard legitimate gendered pairs than to over-include illegitimate gendered pairs (see Section 8 for further discussion).

[5]Due to the filtering process, each language has a slightly different set of sentences for each stereotype in EuroGEST.

datasets, sampling randomly to mirror the distributions of gendered and non-gendered sentences in each 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 judgements (Section C.1). For each sentence, translators also indicate whether the sentence subject is grammatically neutral, masculine or feminine. To explore inter-annotator agreement (IAA), we repeat these annotation tasks for 15 languages, selected to balance language-family diversity with resource and translator availability.

Translation quality The average translation quality across all annotators for the 29 languages was 90.8/100 (Figure 8). However, the Maltese translations were consistently evaluated as lower quality than for other languages. To explore the robustness of the direct assessment scores, for the 15 languages with two sets of annotations we compute Pearson's correlation coefficient (ρ) to measure IAA (Table 4). Average ρ is 0.37, but ρ scores for German, Ukrainian, and Catalan were particularly low. However, when we repeat the validation task a third time for these three languages and exclude outlier annotators, ρ scores are satisfactory (Table 5).

Gender label quality The translators agreed with our system's gendered labels (neutral, masculine or feminine) for 95.9% of sample sentences across all 29 target languages (Figure 8), and 94.5% across only the gendered target languages (for which label assignment is harder). The average Cohen's Kappa score κ between each annotator for the 15 languages with two sets of annotations is 0.81 (near perfect agreement), and annotators disagree with each other on label assignments in only 3% of samples across all 15 languages (Table 6).

3.4 Template Construction

The gender-neutral sentences in EuroGEST$_N$ must be wrapped in a masculine and feminine gendered template (see Table 1) in order to form a minimal pair that can be used to evaluate LLM bias, following Pikuliak et al. (2024). To obtain these templates for all 30 languages, we translate the pronoun and noun-based gendered templates in Table 1 into each language using the Google Translate API. We then validate each template with expert translators by presenting them with a sample genderless sentence wrapped in each of the four gendered templates (Table 1) in that language. Translators rate the four templated sentences on a scale of 0 to 100 (with the same judgement boundaries as for the validation task described in Section 3.3), and are asked to provide a suitable alternative if they give a score of less than 100. The average score was 98.8, and we slightly amend the pronoun-based templates for Catalan, Galician and Italian as a result of the translators' feedback (see Section C.3).[6] Seven languages do not have gendered pronouns or do not use them in these sentence constructions, meaning that, for these languages, only the noun-based template is suitable for creating a gender-minimal pair from a genderless sentence.

4 Method

We use EuroGEST to evaluate generative multilingual LLMs for gender bias by testing the degree to which each LLM prefers stereotypically gendered versions of each sentence in each language (Glaese et al., 2022; Nangia et al., 2020; Pikuliak et al., 2024). We compute the log-likelihoods of the masculine and the feminine version of each EuroGEST sentence S in each model by summing the log probabilities of each token w_t conditioned on the preceding tokens in the sentence during inference (using default model parameters θ). We normalise by the number of tokens T in each sentence to obtain the average log-likelihood $\bar{\ell}(S)$:

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

For each sentence S, we then compute the **relative likelihood of its masculine variant compared to its feminine variant** (r_{masc}). This can be expressed directly using the difference in average log-likelihoods between the two sentences using the logistic sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$:

$$r_{masc} = \sigma\Big(\bar{\ell}(S_{fem}) - \bar{\ell}(S_{masc})\Big). \qquad (2)$$

For each sentence in each language, we therefore obtain an r_{masc} score between 0 and 1 for each sentence in each language, where a score of 0.5 indicates that the LLM attributes equal probability to the two gendered variants. The r_{masc} score is mathematically equivalent to a normalised ratio of the probability of the masculine sentence to the feminine sentence.

[6] We also detail a limitation with the Turkish noun templates in Appendix C.3.

We follow Pikuliak et al. (2024) in defining the **average masculine rate** q_i of each stereotype i as the mean of r_{masc} for all sentences in each stereotype set i.[7] We cannot use these q_i scores directly to measure gender stereotyping, because LLMs exhibit different degrees of default masculine behaviour in different languages, where they tend to prefer masculine forms of words by default because they are more common in the training data and often require fewer tokens. However, the *differences* in q_i scores between feminine stereotypes and masculine stereotypes are indicative of gender stereotyping. We therefore use the q_i scores to measure gender stereotyping in three ways:

1. We use the q_i scores to calculate the **masculine rank** of each stereotype from 1 (most masculine) to 16 (least masculine) in each language, following Pikuliak et al. (2024);
2. We define a **proxy default masculine rate** for each model and language by averaging q_i over seven feminine and seven masculine stereotypes. We then measure the difference between this quasi-neutral baseline and the q_i rate for each stereotype i as an estimation of **inclination** towards i's stereotypical gender[8];
3. We follow Pikuliak et al. (2024) in combining the means of q_i scores for feminine (q_f) and masculine (q_m) stereotypes into an **overall stereotype rate** g_s for each language for each model as follows:

$$g_s = \frac{q_m}{q_f} \tag{3}$$

The g_s score effectively measures how much more likely the LLM is to use masculine gender for stereotypically masculine sentences compared to stereotypically feminine sentences. A g_s score of > 1 or < 1 indicates stereotypical or anti-stereotypical reasoning respectively.

5 Experimental Design

We use EuroGEST to evaluate a range of open-source, pre-trained, decoder-only, Transformer-based multilingual LLMs in order to address our second and third research questions (Section 1). We first consider three LLMs which perform strongly on a range of benchmarks in European languages:

- **EuroLLM** models (Martins et al., 2025) are available in 1.7 billion and 9 billion parameter sizes, both base and instruct. They support the 24 official languages of the European Union (EU) plus eleven 'strategic' languages.
- **Salamandra** (Gonzalez-Agirre et al., 2025) is a suite of base and instruct models with 2, 7, and 40 billion parameters; they support 35 European languages, including all official EU languages and some regional ones.
- **Teuken** (Ali et al., 2025) is an instruction-finetuned 7 billion parameter model which supports all EU languages.[9]

All three families of models use high proportions of non-English training data (between 50 and 60%). We also evaluate three commercial multilingual model families, which do not support as many European languages but which are frequently used for multilingual modelling tasks:

- Alibaba Cloud's **Qwen 2.5** series (Yang et al., 2025) supports more than 30 European and non-European languages, featuring models ranging from 0.5 to 72 billion parameters.
- **Aya Expanse** models, developed by Cohere (Dang et al., 2024), are available with 8 and 32 billion parameters; their strong performance on 23 European and non-European languages is achieved through data arbitrage, multilingual preference training and model merging.
- Meta's **Llama 3** model series (Grattafiori et al., 2024) includes base and instruct models of 1 to 405 billion parameters. They are optimised for 8 languages (6 of which are European) but trained on data including a broader range of languages.

For each model, we calculate r_{masc} for each sentence from each stereotype following Section 4, running inference on each model using their default parameters on NVIDIA-A100 GPUs.

6 Results

Across all languages, all of the LLMs we evaluate consistently show lower average q_i rates for fem-

[7] Unlike Pikuliak et al. (2024), our r_{masc} scores are based on the likelihood of all tokens in each sentence (rather than isolated gendered tokens) and we calculate a normalised ratio of probabilities so that our q_i scores range between 0 and 1.

[8] For example, if the **proxy default masculine rate** in a specific language and model is estimated as 0.6, and the q_i rate for sentences from the *women are neat* stereotype is 0.45, the **inclination** towards the stereotypical gender is 0.15.

[9] At the time we conducted our experiments, only an instruction-finetuned Teuken model is publicly available.

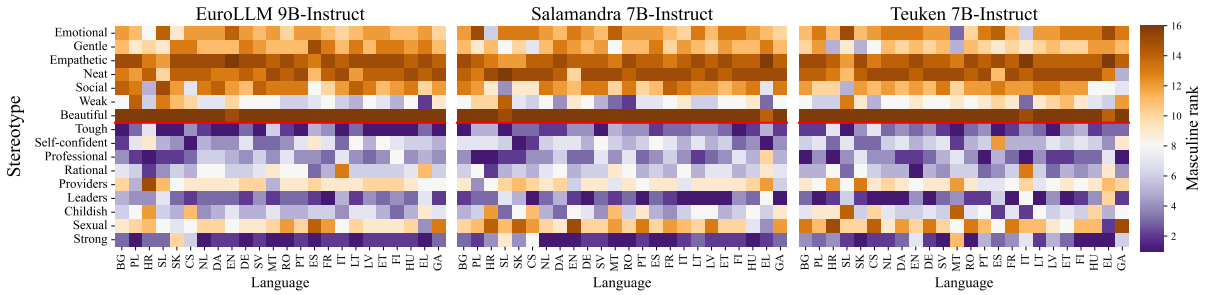

Figure 3: Masculine rank of each stereotype in each official language of the EU in three mid-sized European-centric LLMs. Rank 1 = most strongly associated with masculine gender; Rank 16 = most strongly associated with feminine gender. Red lines divide feminine (top) from masculine (bottom) stereotypes.

inine stereotypes than for masculine stereotypes, regardless of whether we test with noun-based templates (Figure 9), pronoun-based templates (Figure 10) or with morphologically gendered pairs (Figure 11). In this section, we illustrate specific findings relevant to our second and third research questions, using r_{masc} scores with pronoun-based templates for languages with this construction, and noun-based templates otherwise, for evaluating the gender-neutral EuroGEST sentences.

6.1 Patterns of gender stereotyping across languages and LLMs

To address our second research question, we compare which of the 16 gendered stereotypes are most salient across different languages in the three European-centric models (Section 5). To enable a fair comparison across models, we select mid-sized, instruction finetuned versions of each model,[10] and consider only the official languages of the EU (which are supported by all three models).

The results (Figure 3) show the **masculine rank** (see Section 4) of each stereotype in each language in each model. Most masculine stereotypes have clearly higher q_i scores than feminine stereotypes in each language, indicative of stereotypical reasoning. The strongest feminine stereotypes (i.e. the stereotypes with the lowest q_i scores and therefore low masculine ranks) are that women are *beautiful*, *empathetic*, and *neat*, while the strongest masculine ones are that men are *strong*, *leaders*, *tough* (EuroLLM) and *professional* (Salamandra and Teuken). The exceptions are that men are *providers* and *sexual*, and that women are *weak*, where masculine ranks demonstrate neutrality or antistereotypical reasoning in most languages.

Some language-specific outliers are consistent across the three LLMs. For example, in Croatian, the *men are professional* stereotype has particularly high masculine ranks; in Czech the *men are childish* stereotype is always ranked as more feminine than masculine; and in Slovenian the *women are weak* stereotype is firmly feminine-coded across models. Other language-specific results vary by model; for instance, *men are strong* has an unusually low masculine rank in Slovak only in EuroLLM, and *women are emotional* is the most feminine stereotype in Polish only in Salamandra.

6.2 Impact of model size and instruction finetuning on gender stereotyping

To address our second research question, we first explore how a model's size impacts the degree of stereotyping it exhibits. To isolate the impact of model size alone on stereotyping, we first compare five different sizes of Qwen 2.5 models ranging from 0.5 to 14 billion parameters.[11] For each model and each of the 30 EuroGEST languages, we first calculate q_i for each stereotype i and then calculate the **inclination** (see Section 4) of q_i towards the stereotypical gender of i. We average these scores across all languages to obtain an overall measure of the strength of each individual stereotype.

We report results for both base and instruct models in Figure 4. The Qwen 2.5 models clearly reproduce the same strongest and weakest gender stereotypes as the three European-centric models in Figure 3. There is also a visible increase in stereotype strength with model size, as all stereotypes apart from the three weakly encoded ones (*women are weak, men are providers and men are*

[10] `EuroLLM 9B-Instruct`, `Salamandra 7B-Instruct` and `Teuken 7B-Instruct`.

[11] `Qwen 2.5-0.5B`, `Qwen 2.5-1.5B`, `Qwen 2.5-3B`, `Qwen 2.5-7B` and `Qwen 2.5-14B`. We also test the instruct variants of each model.

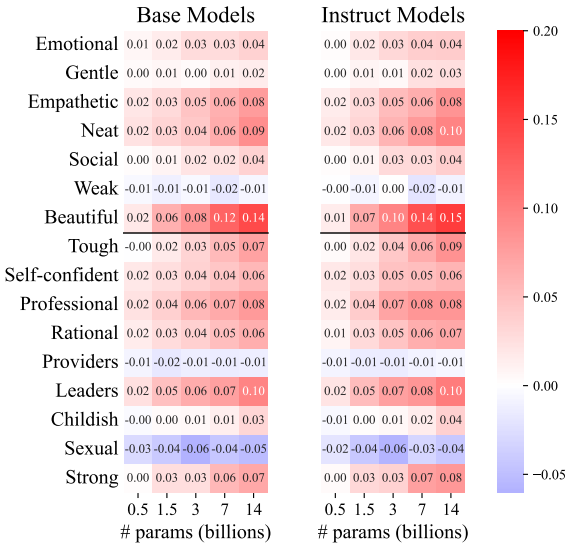

Figure 4: Divergence of q_i scores for each stereotype from proxy default masculine rate towards stereotypical gender for feminine (top) and masculine (bottom) stereotypes in five sizes of Qwen 2.5 models.

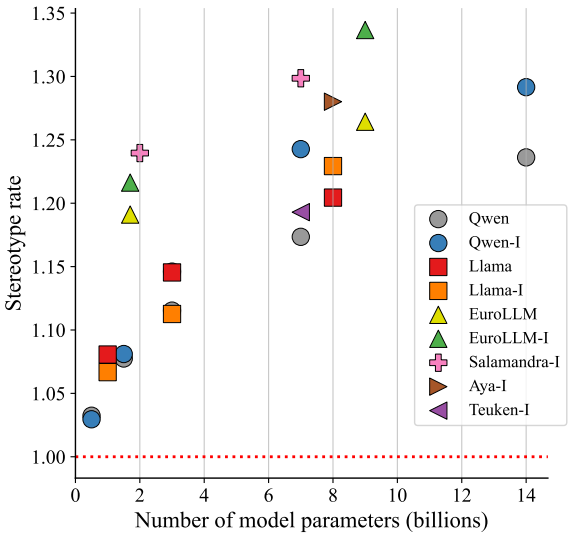

Figure 5: Average stereotype rates of base and instruct models across all languages in EuroGEST. g_s of 1.0 (dotted red line) is indicative of no stereotyping.

sexual) show greater inclination towards stereotypical sentences as model size increases.

Finally, we compute g_s scores for all languages for the 13 models already evaluated and for four smaller EuroLLM and Salamandra models,[12], six Llama models[13] and Aya Expanse.[14] We select these models in order to evaluate a diverse set of multilingual model families and sizes, within the scope of our available resources for inference.

We show the average g_s scores over all languages in Figure 5 (with g_s scores per language shown in Section D). Even the smallest models demonstrate stereotypical reasoning, but g_s increases consistently with model size across all families. Where both base and instruct models are available, instruction-finetuning does not appear to have uniformly decreased gender bias, and in some the instruct models exhibit more stereotypical reasoning than their base model counterparts. We also note that models with broader coverage of EuroGEST languages (such as EuroLLM, Salamandra, Teuken, and Aya) tend to exhibit higher average g_s scores than Qwen and Llama. These high g_s scores likely reflect the fact that these models are more highly performant in general on the full range of European languages, compared to the commercial models.

[12] `EuroLLM 1.7B`, `EuroLLM 1.7B-Instruct`, `EuroLLM 9B`, `Salamandra 2B-Instruct`

[13] `Llama 3.1 1B`, `Llama 3.1 3B` and `Llama 3.1 8B`. We also test the instruct variants of these models.

[14] `Aya Expanse 8B Instruct`

This hypothesis is supported by the per-language g_s scores (Figure 12), which show that all models generally have higher g_s scores on languages which they formally support, compared to those for which they only have latent abilities.

7 Discussion

Our dataset creation process highlights both the promise and the challenges of scaling gender bias evaluation tools across a wide range of languages. It also raises a central methodological question: what constitutes data that is "good enough" for evaluating LLMs for bias? Even in the relatively well-resourced languages represented in EuroGEST, the complexities of morphological gender marking complicate the process of benchmark data translation. However, with the exception of Maltese, our synthetic data generation method was positively evaluated by professional translators in each language, and produces data that we think is sufficiently robust to illustrate systematic gender biases in LLMs across a broader range of languages than has previously been explored. Even with the Maltese dataset of limited quality, we still see evidence of gender stereotyping by the LLMs we evaluate in Maltese (Figures 3 and 12).

Our multilingual bias evaluations also contend with the difficulties of comparing token likelihoods across models and languages, given the fundamentally different distributions of gendered terms and the varied ways in which gender is expressed mor-

phologically, and differences in tokenisation which may impact raw likelihoods of certain terms. By building on the methods developed in Pikuliak et al. (2024) to handle these difficulties, we convincingly show that 13 out of the 16 gendered stereotypes we investigate are consistently present in the internal representations of multilingual LLMs, across all 24 models and 30 languages studied. Portrayals of women as *beautiful*, *empathetic*, and *neat*, and portrayals of men as *leaders*, *strong*, *professional*, and *tough* emerge as the most strongly encoded stereotypes across all models and languages (Figure 7). These findings align with those of Pikuliak et al. (2024), who observed similar patterns across masked, generative, and translation models. We also replicate their observation that *men are sexual* stereotype sentences are more commonly associated with the feminine grammatical gender (Figures 7 and 4), likely reflecting the broader sexualisation of women in text (Pikuliak et al., 2024).

We further find that larger models generally exhibit stronger gender stereotyping (Figures 4 and 5), consistent with prior work (Pikuliak et al., 2024; Tal et al., 2022). This is particularly true in languages where their overall performance is strong. This is intuitive: gender biases emerge as complex distributional patterns in linguistic data, and more powerful models are better equipped to capture them. Models with more multilingual training data will have been exposed to a greater range of multilingual and multicultural expressions of gender stereotypes in their training data. Larger models are also better at modelling morphological and semantic patterns across languages, and we can think of gender stereotyping as complex patterns which can be inducted by LLMs from training data. Importantly, we also observe that instruction-finetuned models often display stronger gender stereotyping than their base-model counterparts. This underscores the unpredictable effects of instruction tuning, which may inadvertently exacerbate harmful representational patterns in some languages, even as it mitigates them in others.

The persistence and salience of these stereotypes in multilingual LLMs may contribute to a range of representational harms as they are deployed in practice, including erasing the visibility of men and women in different roles and contexts and reinforcing discriminatory behaviour and assumptions over time. The subtlety of these biases and the different ways in which they are expressed across languages makes it difficult to evaluate them in practical downstream tasks, yet the need for robust, multilingual evaluation and mitigation grows more urgent as models increase in size and capability. We hope that EuroGEST will provide a foundation for research into how training data, modelling choices and alignment strategies impacts gender bias in multilingual LLMs by offering a resource that enables the systematic evaluation of gender stereotypes in LLMs across languages. Ultimately, consistent cross-lingual mitigation and evaluation strategies will be essential to ensure that increasingly powerful LLMs do not entrench or amplify gendered harms.

8 Conclusion

As LLMs become more powerful and multilingual, it is increasingly important to devise robust evaluation methods to understand how they encode complex social constructs across languages and to minimise the risks of bias and discrimination. With EuroGEST, we extend and release an existing gender bias benchmark dataset (Pikuliak et al., 2024) in 30 European languages. We also document our resource-efficient method for rapidly and sensitively scaling benchmark data across multiple languages. Beyond gender, this approach may also benefit other areas of responsible AI where language coverage remains a critical gap.

Using EuroGEST, we demonstrate that six families of LLMs systematically encode at least 13 gendered stereotypes. We also show that larger and more powerful models exhibit stronger stereotypical biases and reasoning on supported languages and that instruction-finetuned models are sometimes more biased than their base counterparts. These findings highlight the need for mitigation strategies that are both cross-lingual and sensitive to the diversity of gender representations.

This work fills an urgent gap in the lack of multilingual bias evaluation resources. However, its linguistic breadth necessarily limits its depth, and our one-size-fits-all approach certainly cannot capture the full diversity of how gender bias is expressed and experienced by LLM users across all the languages we consider. We hope that our findings will motivate sustained, participatory efforts with gender minorities and experts across diverse linguistic and cultural contexts to develop evaluation methods and resources that move beyond surface-level benchmarking towards more inclusive and socially grounded approaches.

Acknowledgements

This work was supported by the UKRI AI Centre for Doctoral Training in Designing Responsible Natural Language Processing, grant number EP/Y030656/1; EU Horizon Europe Research and Innovation programme grant No 101070631 and UKRI under the UK HE funding grant No 10039436; and the National Science Centre, Poland 2023/49/N/ST6/02691. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

The authors would like to sincerely thank Alessandra Terranova and Lara Dal Molin (University of Edinburgh) for their participation in a pilot version of the validation study and for providing valuable feedback. We are also grateful to Ben Peters (Instituto de Telecomunicações/INESC-ID) for his careful review of EuroGEST's Portuguese samples, which provided insightful observations on potential limitations and systematic patterns in the gender heuristics employed.

Limitations

Scope of biases examined in this work We investigate sixteen specific gendered stereotypes, originally identified in previous work by gender studies experts and literature reviews, by comparing the likelihood of stereotypical and anti-stereotypical sentences during text generation. These are but a small subset of the ways gender bias may arise as LLMs are applied to specific tasks or contexts; future work could 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.

We investigate stereotypes commonly held about men and women, but we do not address LLM biases about people of other genders, as explored in previous work (Blodgett et al., 2020; Dev et al., 2021; Goldfarb-Tarrant et al., 2023; Talat et al., 2022; Munro and Morrison, 2020). This decision is primarily motivated by the lack of standardised gender-inclusive inflection conventions in the gendered languages in our set of target languages. However, we acknowledge that this practical decision risks reinforcing an exclusive or binary understanding of gender, overlooking or minimising ways in which LLM biases may impact non-binary 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.

Finally, we 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.

English-centricity 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 localised 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 behaviour across global populations.

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 be less grammatical or 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 tokenisation should be developed.

Automatic translation We utilise automatic translation for resource-efficient scaling of EuroGEST, and while we employ quality evaluation through both COMET-QE filtering and human validation of a subset of the dataset, we cannot guar-

antee that all EuroGEST sentences are correctly and fluently translated into each language. Automatic translation is not as effective for the lower-resourced languages in the dataset, which reduces both the quality of the translations (Figure 8) and the number of available sentences for evaluating models in these languages (Figure 7).

To identify gendered minimal pairs, we rely on a one-size-fits-all heuristic of less than two letters different on less than one word. This results in both under-inclusion of legitimate gendered pairs where gender is expressed on multiple words or through longer suffixes than allowed by the heuristic, but also over-inclusion of illegitimate pairs. For example, in some cases a random error or variation in translation results in different words which happen to differ by only two letters, even though they are not gendered pairs. Furthermore, for sentences referring to romantic relationships in some languages, this heuristic also captures cases where the object of the sentence (rather than the subject) is gendered in ways which reflect an assumption of heterosexuality by the Google Translate API.[15] While this is certainly an issue we would like to address in future work, it is unfortunately quite likely that the LLMs we are measuring *also* make the same heterosexual assumptions (e.g. interpreting the sentence with feminine object *'artista'* to be indicative of a masculine subject, and masculine object *'artisto'* to be indicative of a feminine subject). The existence of these samples therefore does not necessarily undermine the reliability of the data and the results, depending on how the sentences are used in practice.

Future work could develop language-specific heuristics which more carefully avoid these error cases and retain a higher proportion of the legitimate gendered minimal pairs in each language, depending on the number of words usually gendered in a given sentence or the length of gendered suffixes in that language.

References

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, Katrin Klug, Jasper Schulze Buschhoff, Lena Jurkschat, Hammam Abdelwahab, Benny Jörg Stein, Karl-Heinz Sylla, Pavel Denisov, Nicolo' Brandizzi, Qasid Saleem, and 22 others. 2025. Teuken-7b-base & teuken-7b-instruct: Towards european llms. *arXiv preprint arXiv:2410.03730*.

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.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

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*

[15] For example, the object *'artist'* of the English sentence *"I asked the artist on a date"* is usually translated into languages like Portuguese with feminine gender *'artista'* when the subject is indicated to be masculine, and masculine gender *'artisto'* when the subject is indicated to be feminine, reflecting an assumption of heterosexuality.

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. *arXiv 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*.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 5 others. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

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, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

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. *arXiv preprint arXiv:2409.16235*.

Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Maria Symeonaki, and Giorgos Stamou. 2025. Assumed identities: Quantifying gender bias in machine translation of gender-ambiguous occupational terms. *arXiv preprint arXiv:2503.04372*.

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.

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. *arXiv preprint arXiv:2207.04672*.

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, Stefan Oresko, Andrea Hrckova, and Marian Simko. 2024. Women are beautiful, men are leaders: Gender stereotypes in machine translation and language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3060–3083, Miami, Florida, USA. Association for Computational Linguistics.

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

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

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. 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*.

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.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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. *arXiv preprint arXiv:2309.09120*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

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. *arXiv 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

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.

Dataset statistics per language Figure 6 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. Figure 7 shows the numbers of sentences (both gendered and genderless) remaining for each language, broken down by stereotype category.

C Human validation

Initial evaluation of 100 sentences in 29 languages cost £1,479.00 with a professional translation company, including project management fees. The second round of evaluation of 100 sentences in 15 languages cost £818.55, and the third round of evaluation of 100 sentences in 3 languages cost £163.71. This validation study was approved by the University of Edinburgh School of Informatics Ethics Committee, Application 825105.

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 on each first-person singular sentence in each language in EuroGEST dataset.

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.

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 labelled 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

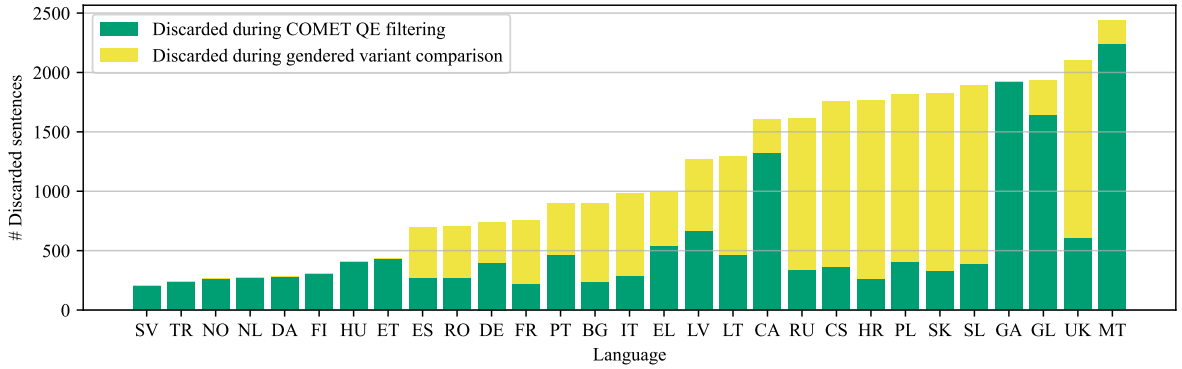

Figure 6: Number of sentences discarded in each language during COMET Quality Estimation filtering or during gendered minimal pair filtering (for gendered languages only).

Stereotype Number

Language	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Bulgarian	195	160	196	146	160	142	183	174	174	161	184	163	170	143	165	146
Catalan	148	118	168	101	127	107	122	115	124	121	150	152	115	98	96	94
Croatian	110	99	133	106	106	81	134	132	116	102	144	109	110	95	112	108
Czech	112	89	144	104	97	75	119	126	124	107	141	126	124	91	119	110
Danish	233	205	239	186	184	188	221	224	211	203	221	208	215	174	188	181
Dutch	234	203	242	190	194	182	222	209	216	199	224	216	217	176	188	182
Estonian	229	197	238	177	187	170	202	196	201	200	221	200	205	165	177	168
English	254	215	256	207	200	197	243	251	229	215	231	222	222	194	208	221
Finnish	238	199	245	184	192	179	221	211	209	204	219	207	211	174	179	185
French	209	167	216	162	156	153	182	189	179	161	198	188	177	150	165	155
Galician	119	98	150	84	115	81	93	101	88	95	120	137	98	78	94	81
German	212	175	208	158	167	159	196	191	181	159	196	182	164	157	169	155
Greek	186	161	192	143	151	143	169	173	169	155	186	174	151	131	152	136
Hungarian	220	196	242	170	192	174	205	206	207	198	219	205	207	171	178	164
Irish	115	111	138	61	117	71	67	91	109	123	138	137	132	80	84	64
Italian	184	165	181	158	152	128	169	181	162	146	188	169	162	136	145	156
Latvian	162	149	187	127	143	121	149	138	133	140	171	161	149	120	132	115
Lithuanian	154	130	179	141	127	111	157	165	144	126	171	147	127	131	134	123
Maltese	80	58	91	27	73	33	43	57	97	102	106	87	103	51	67	46
Norwegian	237	208	237	187	189	180	223	220	212	207	220	214	214	179	191	181
Polish	102	98	129	105	96	71	115	124	119	120	148	118	118	89	101	94
Portuguese	195	154	200	150	158	144	178	175	170	164	188	176	174	144	154	140
Romanian	207	166	224	162	162	152	183	198	181	176	199	189	188	150	165	160
Russian	116	110	144	111	114	77	144	118	119	124	168	126	152	101	116	110
Slovak	117	92	132	102	97	73	115	124	118	108	139	114	108	88	113	98
Slovenian	107	94	125	102	82	77	108	115	117	103	137	114	110	82	98	103
Spanish	213	177	218	167	170	157	189	196	182	176	201	183	172	148	160	156
Swedish	241	209	248	195	194	190	227	229	218	207	225	213	214	179	190	182
Turkish	239	206	245	186	195	174	225	227	216	207	218	214	220	175	189	186
Ukrainian	78	92	110	75	82	56	98	82	89	106	127	111	119	75	83	83

Figure 7: Number of sentences in each stereotype category for each language, across both gendered and genderless EuroGEST datasets.

is lost in the translation.

33: Some meaning preserved: Some of the meaning is preserved but significant parts are missing. 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 by up to three expert translators per language, including both the accuracy ratings via direct assessment and the percentage of gender labels provided by each annotator which align with the label assigned by our translation pipeline. Table 4 shows the Pearson correlation coefficients between the first and second annotators' direct assessment scores (for the 15 languages for which we have two sets of annotations). Table 5 shows the Pearson correlation coefficients between two annotators excluding an outlier annotator (for the 3 languages for which we have three sets of annotations). Table 6 shows the Cohen's Kappa scores between the first and the second annotators' gender labels for the 100 sample sentences for each language for which we have two annotators. We note that for Estonian and Finnish, Cohen's Kappa score is not calculable because there is no variation between the two sets of gender labels (all are genderless sentences). We expected the same for Dutch and Swedish, which are also genderless languages, but we observed that in a very small number of cases the annotators labelled sample sentences in these languages as grammatically gendered. This was due to the presence of specific gendered nouns which are a vestige of Dutch's grammatical gender system, e.g. *vrachtwagenchauffeur/vrachtwagenchauffeuse* or *vriend/vriendin*. The nature of Cohen's Kappa scoring means that where most labels are the same category, disagreements on non-majority category labels like this are more heavily penalised, hence the relatively low Kappa scores for Dutch and Swedish.

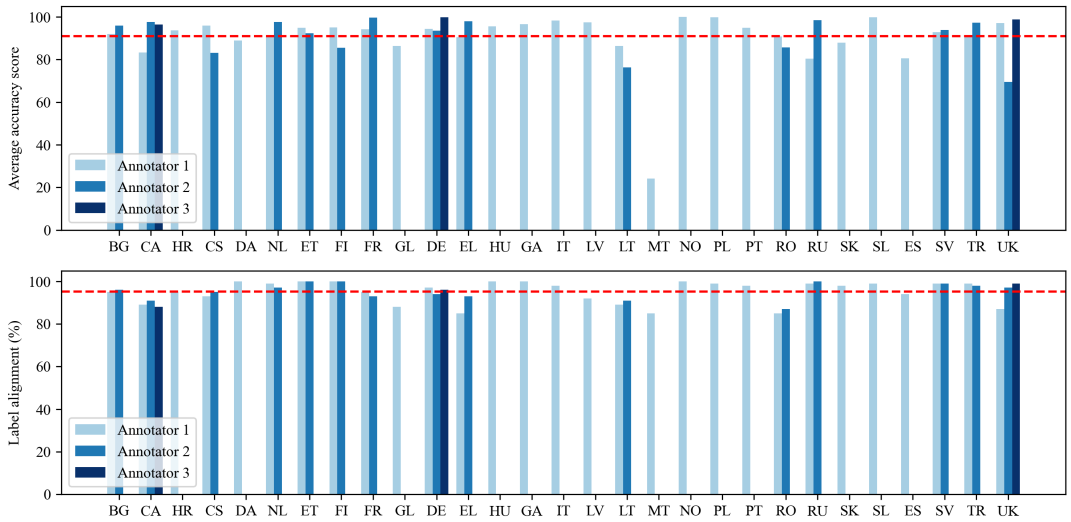

Figure 8: Average ratings for EuroGEST sentence translation quality (top) and gender label accuracy (bottom) for sample of 100 GEST sentences in each language, with up to three annotators per language.

Language	Pearson ρ	Pearson p-value
Bulgarian	0.3880	0.0001
Catalan	0.1246	0.2190
Czech	0.4306	0.0000
Dutch	0.2782	0.0051
Estonian	0.2240	0.0251
Finnish	0.5995	0.0000
French	0.4069	0.0000
German	-0.0150	0.8820
Greek	0.7517	0.0000
Lithuanian	0.5897	0.0000
Romanian	0.5897	0.0000
Russian	0.3014	0.0023
Swedish	0.6265	0.0000
Turkish	0.1863	0.0635
Ukrainian	0.0709	0.4835

Table 4: Pearson correlation and p-values by language for direct assessment scores on sample of 100 sentences per language for the 15 languages with two sets of annotations.

Language	Pearson ρ	Pearson p-value
Catalan	0.1814	0.0723
German	0.1129	0.2635
Ukrainian	0.3595	0.0002

Table 5: Pearson correlation and p-values by language for direct assessment scores on sample of 100 sentences per language for the three languages with three sets of annotations, where outlier annotators are excluded.

C.3 Validated prompt templates in each language

Table 7 shows the masculine and feminine noun and pronoun templates in each EuroGEST language. For Catalan, the *'he said'* template was automatically translated as *'va dir'*, but our translator said that *'va dir ell'* is more appropriate. Conversely, for Galician the *'she said'* template was automatically translated as *'dixo ela'* but the translator corrected this to simply gender-neutral *'dixo'*. For Italian, both 'he said' and 'she said' templates were automatically translated as 'disse', but the translator amended these templates to be 'disse lui' and 'disse lei'.

Finally, for Turkish the translator advised that "'S", the man/woman said' is better translated as 'Adam/Kadın, "S" dedi' than "'S" dedi adam/kadın. However, this sentence-initial noun template is in-

Language	Cohen's κ	# different labels
Bulgarian	0.95022	3
Catalan	0.87664	4
Czech	0.88131	6
Dutch	0.49367	2
Estonian	-	0
Finnish	-	0
French	0.92959	3
German	0.75610	3
Greek	0.75850	12
Lithuanian	0.96151	2
Romanian	0.94848	2
Russian	0.98398	1
Swedish	0.49749	1
Turkish	0.66216	1
Ukrainian	0.79977	12

Table 6: Cohen's Kappa scores by language for gender labels (masculine, feminine or neuter) assigned to 100 sample sentences per language for the 15 languages with two sets of annotations. Number of instances where annotators disagree also displayed for clarity.

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	"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», — сказала 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», — сказала вона	«S», — сказав чоловік	«S», — сказав жінка

Table 7: Gendered noun and pronoun templates in all languages in this work, as validated by expert translators. 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.

consistent the sentence-final template constructions in the other 29 languages, so we do not implement this suggestion (but note that our templated sentences for Turkish may therefore be less grammatical and the results less reliable).

D Additional results

Figures 9, 10 and 11 show the average masculine rates (q_i) on all sentences from feminine stereotypes and all sentences from masculine stereotypes using noun-based templates, pronoun-based templates and gendered minimal pairs respectively, for a selection of six medium-sized models. Figure 12 shows how the g_s rate increases with larger model sizes, displaying the results from all 24 models from six language families on each language.

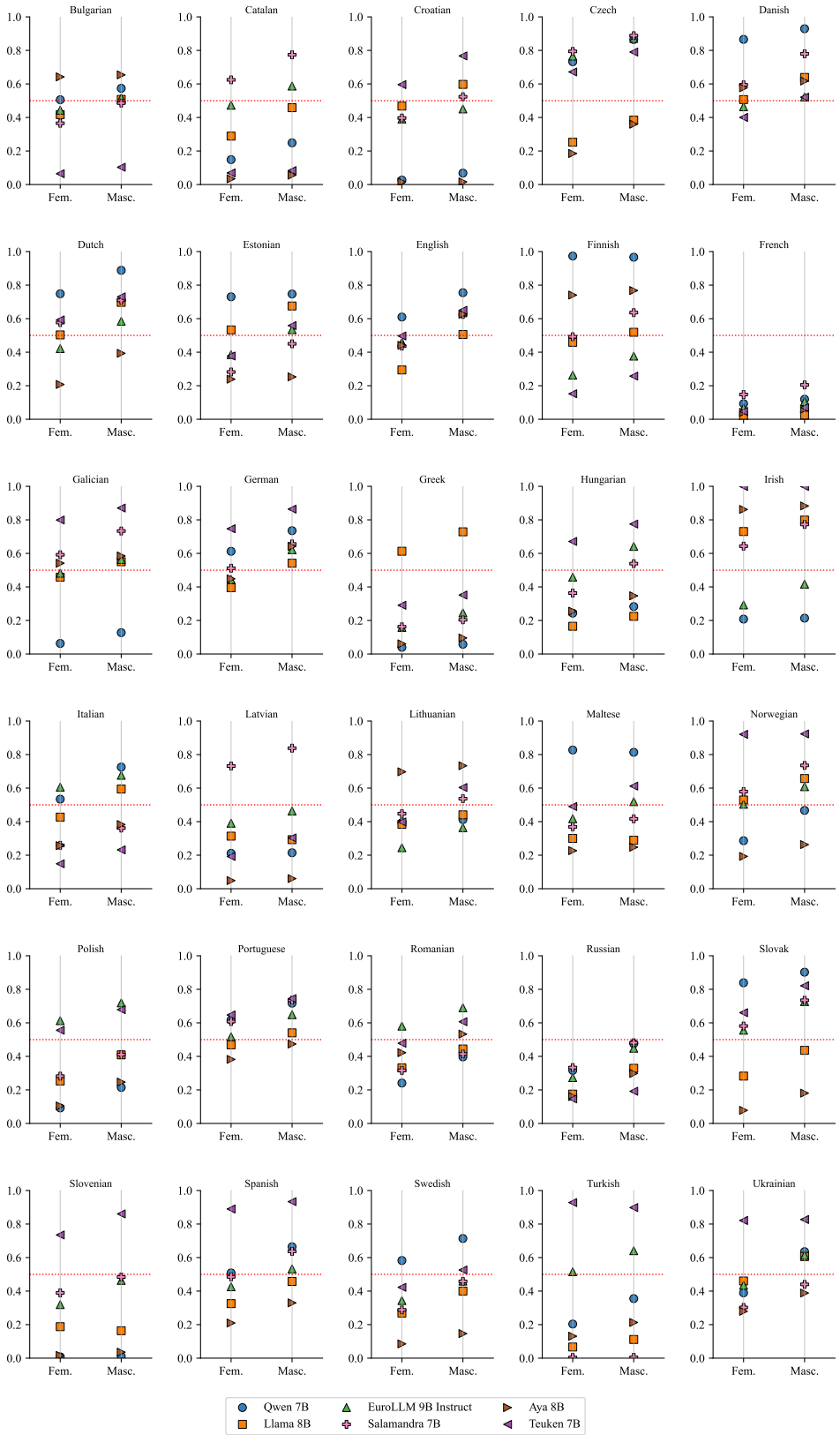

Figure 9: Average q_i rates of six mid-sized models on sentences from all feminine and all masculine stereotypes across all available gender-neutral sentences per language, wrapped in a gendered noun-based template (*"'S,' the man/woman said"*.)

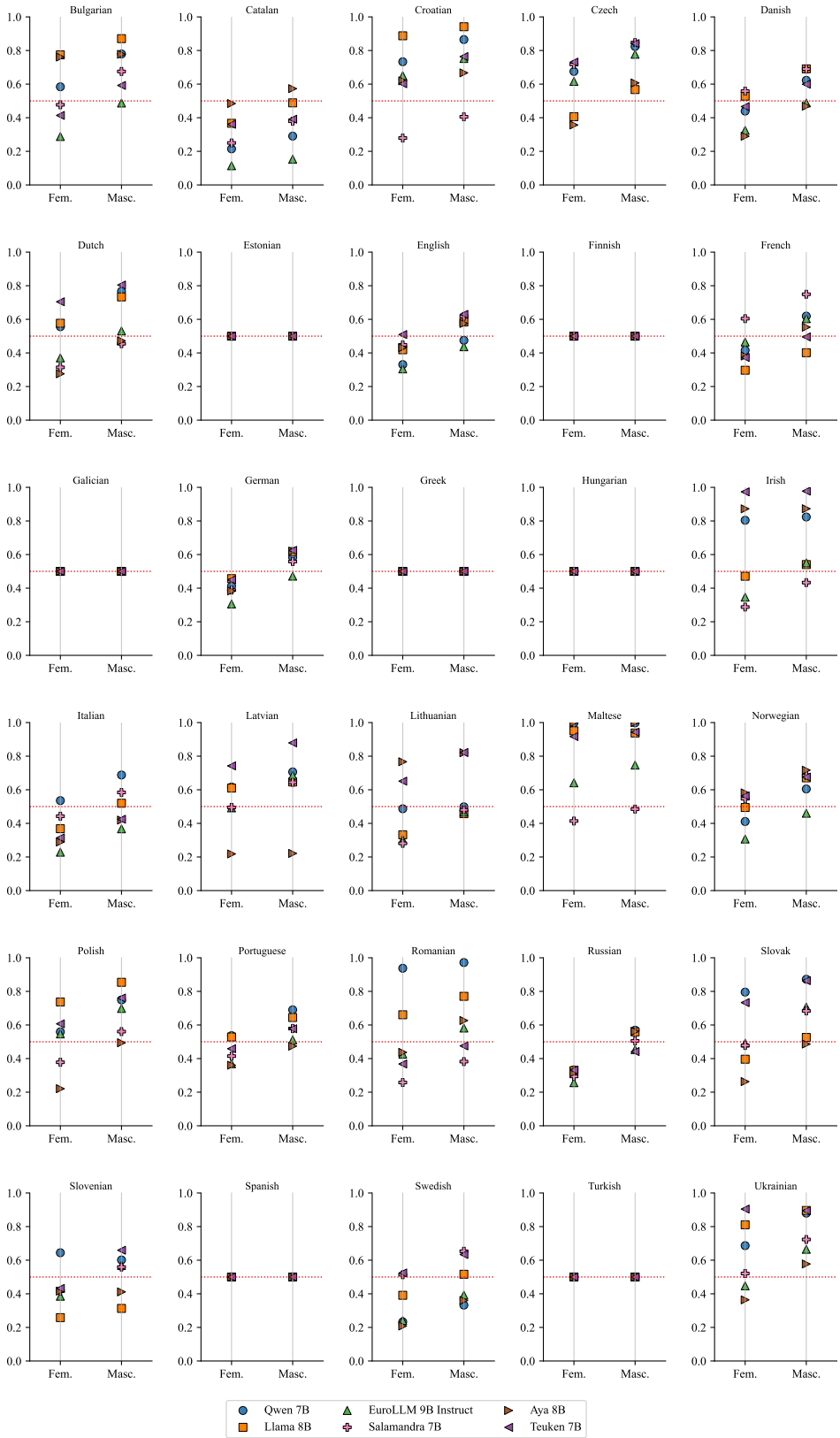

Figure 10: Average q_i rates of six mid-sized models on sentences from all feminine and all masculine stereotypes across all available gender-neutral sentences per language, wrapped in a gendered pronoun-based template (*"'S,' he/she said"*.)

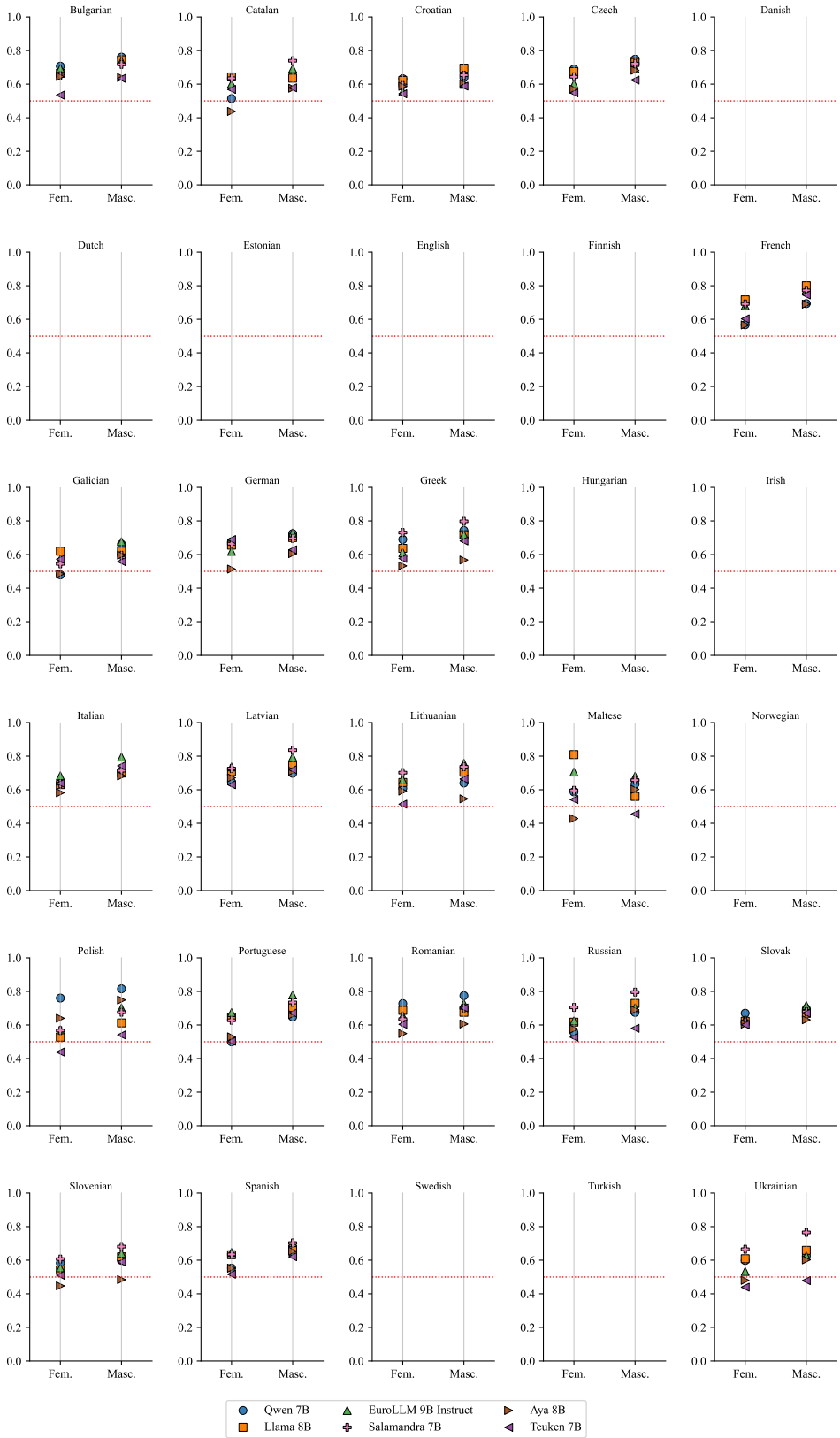

Figure 11: Average q_i rates of six mid-sized models on sentences from all feminine and all masculine stereotypes across all available gendered sentences per language, for languages which mark grammatical gender on some EuroGEST sentences.

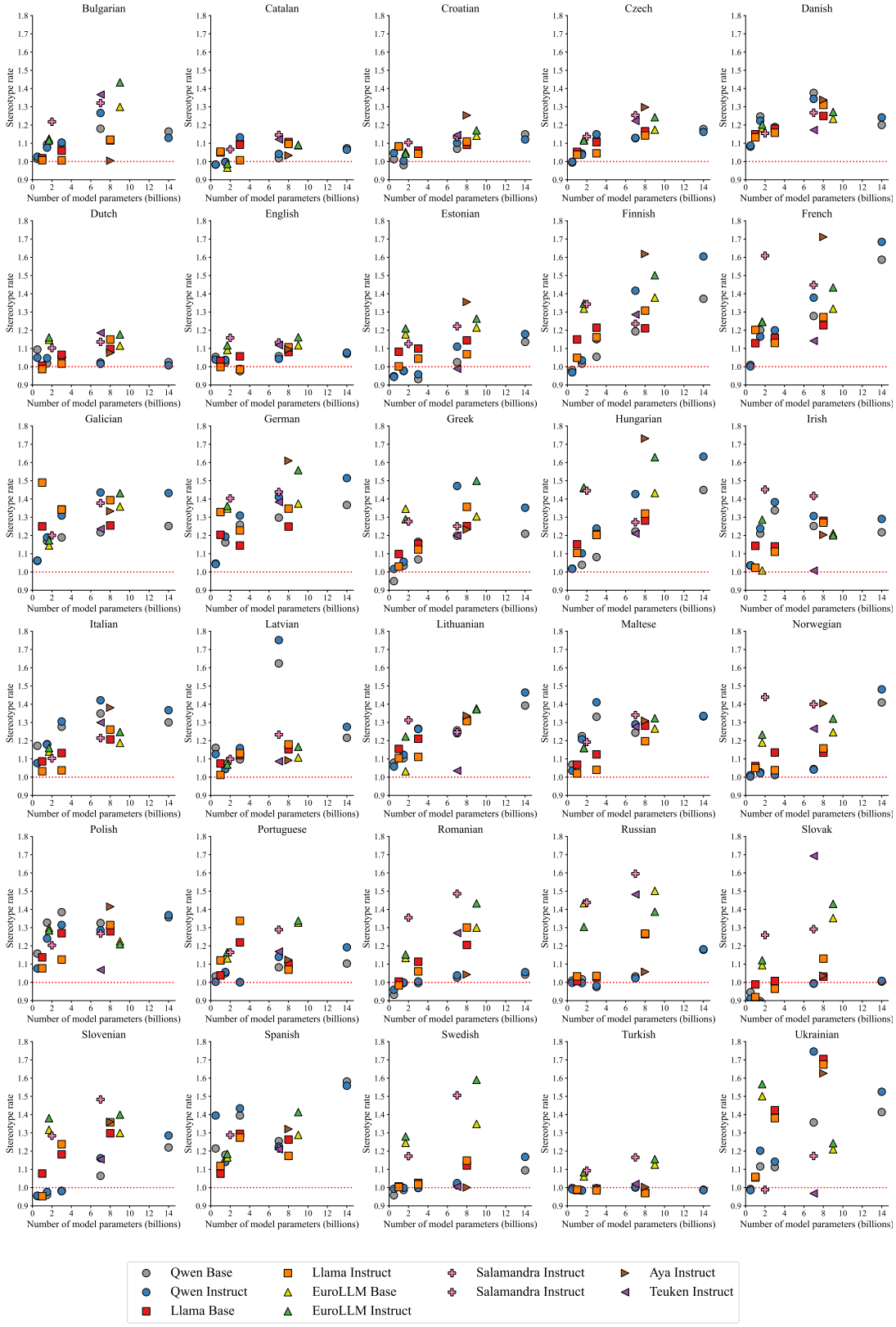

Figure 12: Average stereotype rates of base and instruct models in each language. Stereotype rate of 1.0 is indicative of no stereotyping.