Gender Bias in MT for a Genderless Language: New Benchmarks for Basque

Amaia Murillo, Olatz Perez-de-Viñaspre, Naiara Perez
HiTZ Center - Ixa, University of the Basque Country UPV/EHU
{name.surname}@ehu.eus

Abstract

Large language models (LLMs) and machine translation (MT) systems are increasingly used in our daily lives, but their outputs can reproduce gender bias present in the training data. Most resources for evaluating such biases are designed for English and reflect its sociocultural context, which limits their applicability to other languages. This work addresses this gap by introducing two new datasets to evaluate gender bias in translations involving Basque, a low-resource and genderless language. WINOMT$_{eus}$ adapts the WinoMT benchmark to examine how gender-neutral Basque occupations are translated into gendered languages such as Spanish and French. FLORES+Gender, in turn, extends the FLORES+ benchmark to assess whether translation quality varies when translating from gendered languages (Spanish and English) into Basque depending on the gender of the referent. We evaluate several general-purpose LLMs and open and proprietary MT systems. The results reveal a systematic preference for masculine forms and, in some models, a slightly higher quality for masculine referents. Overall, these findings show that gender bias is still deeply rooted in these models, and highlight the need to develop evaluation methods that consider both linguistic features and cultural context.

Keywords: Bias, Gender, Machine Translation, Less-Resourced Languages, Basque

1. Introduction

Language technologies are often trained on huge amounts of uncurated data extracted from the Internet and, as a result, they can inherit stereotypes, misrepresentations, derogatory language and other demeaning behaviours that disproportionately affect vulnerable and marginalized communities (Gallegos et al., 2024). When deployed in real-world applications, these systems can reproduce and even amplify such biases, leading to harmful social consequences (Salinas et al., 2023). For that reason, evaluating these biases is a key step toward developing fairer models.

Among the different forms of bias, gender bias has been one of the most extensively studied (Devinney et al., 2022; Cignarella et al., 2025). However, most evaluation datasets have been developed in English and reflect the sociocultural context of English-speaking communities (Zulaika and Saralegi, 2025). Since bias is shaped by culture, these resources cannot be directly applied to other languages, where stereotypes and social dynamics differ (Jin et al., 2024; Oppong et al., 2025). Furthermore, these datasets rely on English-specific linguistic features and on cues that may not be applicable to typologically different languages. In fact, this form of bias is often evaluated through explicit gender markers—such as pronouns (Rudinger et al., 2018) or other marked forms (Stanovsky et al., 2019)—but this approach cannot be used in the same manner for languages that lack grammatical gender (Corral and Saralegi, 2022). As a consequence, many languages remain underrepresented in bias research, and Basque, in particular, faces a scarcity of resources for evaluating gender bias, as a low-resource and genderless language.

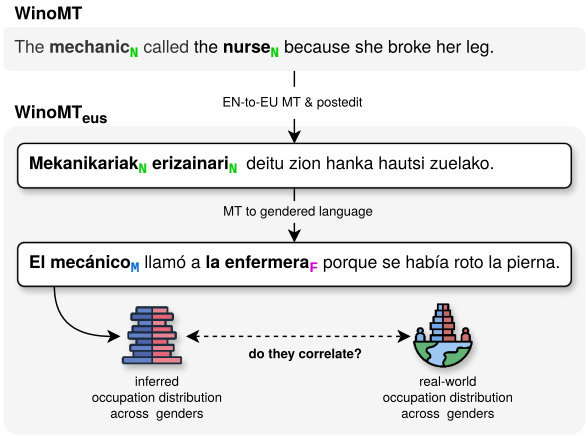

Figure 1: Overview of WINOMT$_{eus}$. Sentences with gender-neutral (N) occupation mentions, taken from WinoMT (Stanovsky et al., 2019) and translated to Basque, are translated into a gendered language; we compare the resulting gendered occupation distribution (male M or female F) with real-world data.

To assess this type of bias in languages like Basque, machine translation (MT) offers a particularly useful framework. By translating between gendered and genderless languages, it is possible to observe how models handle gender information when explicit grammatical cues are present or absent (Zahraei and Emami, 2025; Rarrick et al., 2024). In short, translation not only reveals how

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8971–8984
11–16 May 2026.

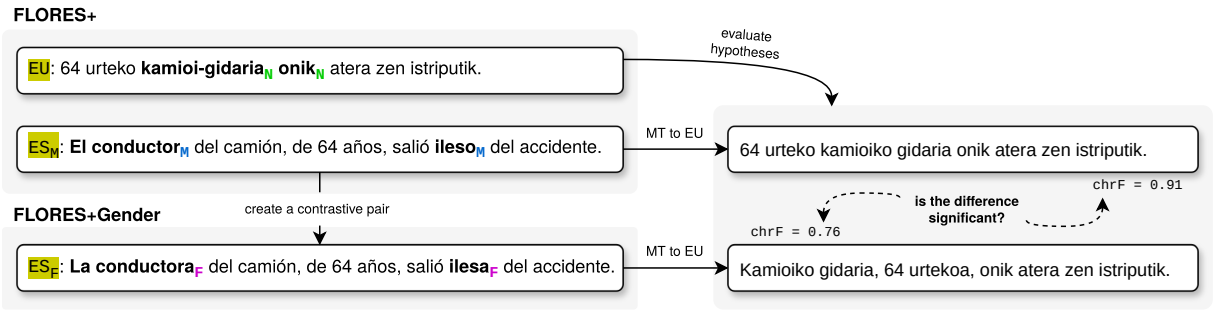

Figure 2: Overview of FLORES+Gender. Using Basque-gendered language pairs from FLORES+,[3] we create a contrastive version of the gendered sentence (male M or female F). Both are translated into Basque, and translation quality is compared to asses the effect of source-sentence gender.

gender disparities can manifest in practical scenarios, but also serves as a tool to uncover systematic gender preferences in the models.

In this work, we release two datasets designed to assess gender-related bias in Basque systems through MT. **WinoMT$_{eus}$**[1] (see Fig. 1) focuses on how gender-neutral Basque occupations are rendered in gendered target languages, and, for the first time, whether they align with official employment statistics from the Basque Country. **FLORES+Gender**[2] (see Fig. 2) studies whether translation quality varies when translating from gendered languages (Spanish and English) into Basque depending on the gender of the referent. Together, these resources provide a framework for evaluating gender bias in Basque MT and contribute to broadening bias evaluation beyond English.

2. Background and Related Work

Gender bias in natural language processing (NLP) is a specific form of social bias that refers to unequal treatment or outcomes based on gender, often caused by historical and structural inequalities. It can manifest as intrinsic bias—encoded in the model's internal representations—or as extrinsic bias—which emerges in downstream tasks such as question answering or sentiment analysis (Guo et al., 2024). These inequities can result in representational and allocational harms, shaping how groups are depicted and how resources are distributed (Crawford, 2017). Evaluating these biases is crucial for building fairer language technologies.

2.1. Bias Evaluation Resources

Most existing bias evaluation datasets to date have been created for English. These include Winostyle counterfactual benchmarks such as Winogender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018), where pairs of minimally different sentences reveal stereotypical associations in model predictions, as well as other sentence-pair resources such as GAP (Webster et al., 2018), BUG (Levy et al., 2021) or BEC-Pro (Bartl et al., 2020). These resources have proved influential but remain tightly coupled to the grammatical and sociocultural properties of English, which features gendered pronouns and agreement patterns that do not generalize to typologically different languages. Prompt-based datasets (e.g., RealToxicityPrompts (Gehman et al., 2020), BOLD (Dhamala et al., 2021)) have extended bias assessment to generative settings, yet such resources are again largely English-specific. Consequently, most low-resource and genderless languages lack dedicated benchmarks for gender bias evaluation.

2.2. Basque and Existing Work

Basque is a language isolate spoken by approximately three-quarters of a million people, most of whom reside in the Basque Country, a territory spanning areas of both Spain and France (Soraluze et al., 2016). Basque does not have grammatical gender, unlike its neighboring languages Spanish and French, and even personal pronouns are gender-neutral. Still, the absence of grammatical gender does not prevent the reproduction of bias, as language use itself can encode and transmit social stereotypes. Gender is conveyed through lexical choices: gender identity terms (*mutil* 'boy', *emakume* 'woman'), kinship terminology (*izeba* 'aunt', *aitona* 'grandfather'), occupations (*andereño* 'female schoolteacher', *maisu* 'male schoolteacher') and proper names (*Aitor, Lander* and *Asier* are typically associated with men, whereas *Ane, Lide* and *Nahia* are commonly given to women).

[1] https://github.com/amaiamurillo/winomteus

[2] https://github.com/amaiamurillo/flores_plus_gender

[3] https://huggingface.co/datasets/openlanguagedata/flores_plus

Despite growing interest, resources for gender-bias evaluation in Basque remains scarce. The most closely related effort is the TANDO project (Gete et al., 2022, 2025), which introduced contrastive Basque-Spanish test sets to examine gender selection and register phenomena in MT. In their setup, Basque source sentences were gender-neutral but required gendered translation to Spanish; translation quality was then compared across masculine and feminine references. TANDO therefore provided the first resource for assessing gender bias in Basque MT, specifically for the Basque-to-Spanish direction.

2.3. Gender Bias Evaluation through MT

MT offers a natural framework for evaluating gender bias in languages without grammatical gender. When translating from a genderless language (e.g., Basque or Turkish) into a gendered one (e.g., Spanish or English), models must make an explicit gender choice. Over many examples, a systematic preference for one gender can reveal underlying bias (Stanovsky et al., 2019; Mastromichalakis et al., 2025). Conversely, translation from a gendered into genderless language allows researchers to measure whether translation quality differs by gendered input forms (Costa-jussà et al., 2023).

Building on this tradition, our work expands the scope of Basque MT bias evaluation in two ways. First, we present **WinoMT$_{eus}$**, a counterfactual Wino-style dataset for translating *from* Basque into gendered languages, where both masculine and feminine translations are plausible, and compare their distribution to official employment statistics, following Mastromichalakis et al. (2025). Second, we introduce **FLORES+Gender**, which extends FLORES+ to evaluate translations *into* Basque from gendered languages (Spanish and English), testing whether gendered source forms affect translation quality. TANDO included Spanish to Basque but not English to Basque, making this the first multilingual setup of its kind for Basque.

3. WinoMT$_{eus}$

We translated and adapted WinoMT (Stanovsky et al., 2019) into Basque to examine how gender is assigned when translating gender-neutral occupations into gendered languages. The original dataset is a concatenation of Winogender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018), which are two widely known template-based datasets for assessing occupational stereotypes through coreference resolution. Each instance includes two gender-neutral occupations and a pronoun that refers to one of them, and the model is required to resolve the ambiguity (see Fig. 1).

This task cannot be applied to Basque in the same way as in English, given that Basque lacks gendered pronouns. Hence, we translated the original dataset and created a Basque version, called WinoMT$_{eus}$, to analyze how models render these professions when translating into gendered languages such as Spanish or French. The adaptation process involved five main stages:

1. **Glossary creation.** We created a glossary of 78 occupations to align English terms with culturally and linguistically equivalent forms in Basque. This process required dealing with issues such as the lack of lexical equivalents or professions without cultural counterparts.
2. **Translation.** The original English dataset was translated into Basque using GPT-4o.
3. **Post-editing and revision.** We post-edited the translated sentences while addressing the pitfalls identified by Blodgett et al. (2021) present in the original dataset, including grammatical errors, sentence structure issues, logical failures and unnatural phrasing.
4. **Cultural adaptation.** Contextually sensitive elements were adjusted to fit the Basque context (e.g., emergency numbers, currency and units of measurement).
5. **Final filtering.** Duplicated sentences were removed, as several English sentences differ only in pronoun gender and therefore resulted in identical translations in Basque.

The resulting dataset includes **1,827 sentences** in Basque with at least one occupation mention.

4. FLORES+Gender

FLORES+Gender is built on the FLORES+ benchmark (NLLB Team et al., 2024), developed by Meta to assess MT systems for low-resource languages. The original dataset contains 1,012 sentences in English, evenly sampled from Wikinews, Wikijunior and Wikivoyage, translated into around 200 languages and varieties, including Basque, Spanish, Catalan, Valencian and Galician.

Adopting the approach of Costa-jussà et al. (2023), we reversed the translation direction to study whether the grammatical gender of the source text affects translation performance *into* Basque. Specifically, we selected two source languages with different degrees of gender marking: Spanish, a strongly gendered language, and English, a weakly gendered one where gender is mostly conveyed through personal pronouns, possessive adjectives and gender-specific nouns.

For each source language, we construct contrastive versions of the dataset: one containing all the sentences in masculine form and another with the same sentences in feminine form. From the

	ME	PN	UM
Spanish	24.52	19.65	60.01
English	29.03	49.68	n/a

Table 1: Quantification of FLORES+Gender annotated phenomena (% over total sentences).

original **Spanish** set, we identified **363 sentences** with gendered references, and **155** in the **English** version. These were manually adapted to produce gender-controlled pairs while maintaining semantic equivalence. In addition, each sentence was manually annotated for

- **ME**: whether it contains mentions to multiple gendered human entities,
- **PN**: the presence of proper names, and
- **UM**: the presence of unmarked usage of the masculine form (applies only in Spanish).

These annotations, quantified in Tab. 1, support downstream analysis of whether such factors influence translation quality.

When adapting the data, we followed consistent criteria to ensure coherence and cultural plausibility. Proper names were replaced with alternatives of the opposite gender, prioritizing equivalents with similar spelling and cultural origin (e.g., *Cristina* $_{\text{F}}$/*Cristian* $_{\text{M}}$). In the case of highly recognizable names (e.g., *Fernando Alonso*), we employed plausible alternatives whenever the context allows. For multi-entity sentences, we focused on modifying the main subject or the most contextually relevant gendered entity in the sentence instead. We also addressed various errors encountered during the annotation process. These included typos, unnecessary or missing words, as well as translation inconsistencies and gender agreement issues present in the original sets.

Finally, after producing the two gender-specific sets for each source language, we manually generated the two corresponding Basque sets with the same modifications, aimed at serving as a reference for automatic evaluation.

5. Experimental Setup

The goal of our experiments is to assess how gender bias manifests in MT involving Basque, using the two new resources introduced in this work. That is, we design complementary evaluations that examine gender bias both *from* Basque into gendered target languages and *into* Basque from other languages, using controlled datasets and quantitative metrics.

5.1. Evaluation Methodology

This section describes the evaluation procedures employed to analyse gender bias in translation in both directions. We leverage the two datasets presented in §3 and §4 to perform, respectively, *i)* an analysis of how gender is expressed when translating *from* Basque into gendered languages (Spanish and French), and *ii)* an analysis of how translation *into* Basque behaves with respect to gender-marked referents.

5.1.1. WinoMT$_{\text{eus}}$: *from* Basque

This experiment analyses whether translation from Basque into a gendered language reinforces gender stereotypes or reflects actual labour distribution in the Basque Country. The evaluation consists in (see Fig. 1) *i)* machine-translating the dataset into Spanish and French, *ii)* extracting the mentions of occupations in the translations and labelling their gender, and *iii)* comparing the translated occupations' distribution to real-world labour statistics. In what follows, we detail the methodology employed to extract, gender-label, and evaluate the occupations obtained from the translated corpus.

Occupation extraction. We automatically extracted the professions with Claude 4 Sonnet. For the prompt, we adopted a one-shot approach, giving the model an instruction to extract the occupations from the Spanish and French translations, along with one example showing how to perform the task. In this process, we retrieved both the profession and the preceding article, and if no occupational term was detected, the output was marked as `[MISSING]`. To validate this extraction method, we manually checked 100 randomly selected sentences, and obtained 95.5% of accuracy in occupation extraction. Next, we inferred the gender of each profession based on the gender of the article. In cases where the article was not enough to determine the gender, we relied on the morphological ending of the occupation (e.g., *l'infirmier* $_{\text{M}}$/*l'infirmière* $_{\text{F}}$ or *l'administrateur* $_{\text{M}}$/*l'administratrice* $_{\text{F}}$).

Labour statistics. To evaluate the extent to which model biases reflect real-world disparities, we follow prior work by Mastromichalakis et al. (2025), Ibaraki et al. (2024) and Touileb et al. (2022), comparing our results with actual labour distribution statistics. These statistics for the Basque Country were retrieved from Lanbide,[3] the public employment service, as of July 2025. We manually mapped the profession categories provided by

[3] https://www.lanbide.euskadi.eus

Lanbide to our glossary (see §3) in order to identify the closest corresponding occupations. This allowed us to conduct two complementary evaluations: one at the aggregate level, comparing the overall distribution of masculine and feminine translations with labour statistics, and another at the profession level, analysing correlations for each occupation individually.

Evaluation Metrics. On the one hand, we computed the **Pearson correlation coefficient** between the percentages of feminine and masculine translations for each occupation and the corresponding labour statistics to evaluate how closely the gender proportions in the model translations align with real-world data. On the other hand, we applied the **GRAPE** (*Gender RAtion Probabilitiés*) metric, introduced by Mastromichalakis et al. (2025), to carry out a more detailed analysis at profession level. This metric measures gender bias by comparing how frequently a model generates masculine or feminine forms relative to a reference distribution. It quantifies both the direction of the bias (GRAPE-M for masculine, GRAPE-F for feminine) and its magnitude (how strong that tendency is), starting from -1. For instance, a GRAPE-F value of 0.0 means that the system generates the same proportion of feminine forms as the reference distribution, whereas a value of 1.0 means that feminine forms are produced twice as often as expected.

5.1.2. FLORES+Gender: *into* Basque

Adopting the methodology of Costa-jussà et al. (2023), we reversed the translation direction to analyse whether the translation quality varies depending on the gender of the referent in the source sentence. For this purpose, we translated into Basque the English and Spanish versions of the FLORES+Gender dataset. We first evaluated overall translation quality across masculine and feminine subsets to detect global trends. Subsequently, we exploited the dataset's fine-grained annotations to disaggregate results by the linguistic and contextual factors explained in §4 and Tab. 1.

Evaluation metrics. We compared the translations to their corresponding references employing standard automatic metrics using the SacreBLEU toolkit (Post, 2018): character-level F-score (chrF++; Popović, 2017) and Translation Edit Rate (TER; Snover et al., 2006). To determine whether the observed differences between masculine and feminine subsets were statistically significant, we applied the paired approximate randomization test (Riezler and Maxwell, 2005) with 10,000 random permutations and significance levels $p < 0.01$, $p < 0.05$, and $p < 0.1$.

5.2. Models and Inference Setup

With the above described datasets, we assessed how gender bias manifests in MT across three different technical paradigms.

General-purpose LLMs. In this category, we evaluated **Latxa 3.1 8B Instruct**[4] and **Latxa 3.1 70B Instruct**[5], two general-purpose instructed LLMs for Basque developed by Sainz et al. (2025). Both are based on **Llama 3.1 8B Instruct**[6] and **Llama 3.1 70B Instruct**[7] (Grattafiori et al., 2024), which were also included in the evaluation. All four models were deployed with vLLM (Kwon et al., 2023), using a single A100 GPU for the 8B variants and two GPUs for 70B ones. We used the models' default generation settings with the temperature fixed at `0` to ensure reproducible results.

Additionally, we also assessed **GPT-5**[8], accessed via OpenAI API, with reasoning effort set to `minimal`. **Claude Sonnet 4**[9] was evaluated using Anthropic's API with the temperature also fixed at `0`. The last model in this category, **DeepSeek-V3.2-Exp Non-thinking** (DeepSeek-AI et al., 2025), was run via its API using default generation parameters.

Beyond general-purpose models, we also included **SalamandraTA 7B Instruct**[10] (Gilabert et al., 2025), an instruction-tuned translation model derived from SalamandraTA 7B Base that supports 35 European languages, among them Basque, Spanish, English, French, Aragonese, Asturian and Catalan.

Open NMT models. Among open neural machine translation (NMT) models, we tested **MADLAD-400-3B-MT**[11] (Kudugunta et al., 2023), a multilingual model covering 400 languages; **NLLB-200's 3.3B variant**[12] (NLLB Team et al., 2024), designed to support low-resource languages and supports 200 language pairs; and **HiTZ Center's**

[4] https://huggingface.co/HiTZ/Latxa-Llama-3.1-8B-Instruct
[5] https://huggingface.co/HiTZ/Latxa-Llama-3.1-70B-Instruct
[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[7] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
[8] https://cdn.openai.com/gpt-5-system-card.pdf
[9] https://www.anthropic.com/news/claude-4
[10] https://huggingface.co/BSC-LT/salamandra-7b-instruct
[11] https://huggingface.co/google/madlad400-3b-mt
[12] https://huggingface.co/facebook/nllb-200-3.3B

Spanish-Basque and English-Basque MT models[13]. To translate the sentences, we loaded each model and its corresponding tokenizer from Hugging Face using the Transformers library and generated outputs with the default settings.

Proprietary translation services. As a reference for deployed MT systems, we assessed various proprietary translation services. Such systems provide limited transparency with respect to their internal architectures. We included **Google Translate**, the widely used commercial MT service developed by Google that covers over 100 languages (accessed through the Cloud Translation API); **Elia**[14] developed by Elhuyar Foundation and supporting Basque, Spanish, French, English, Catalan and Galician; **Batua**[15], created by Vicomtech and covering Basque, Spanish, French and English; and **Itzuli**[16], provided by the Basque Government and supporting the same languages as Batua. Translations with Elia, Batua, and Itzuli were obtained manually through their public web interfaces.

6. Results

Moderate alignment with labour distributions. This analysis of the WINOMT$_{eus}$ experiments assesses the extent to which model outputs reflect the actual occupational gender distribution. When translating from Basque into Spanish, **GPT 5**, **NLLB-200-3.3B** and **Latxa 3.1 70B Instruct** show the highest correlation ($r > 0.4$, $p < 0.01$), which indicates a moderate positive relationship between the gendered forms produced by the models and the real-world gender distribution of professions (see Tab. 2). For Basque–French translations, correlations are generally lower, although the ranking of systems remains broadly similar. This drop may be due to cross-linguistic differences in gender marking or to disparities in the training data available for French. A closer look at the Llama-based models reveals that **Latxa**'s adaptation to Basque notably improves alignment compared to its **Llama 3.1** counterparts of the same size. This suggests that targeted language adaptation can enhance a model's sensitivity to gender patterns present in specific linguistic contexts. More broadly, models explicitly trained or fine-tuned for translation tasks (e.g., NLLB-200-3.3B, SalamandraTA 7B Instruct, GPT 5) tend to achieve higher correlations than purely general-purpose systems, while smaller instruction-tuned LLMs such as Latxa 3.1 8B and Llama 3.1 8B show weaker alignment.

[13]No Basque-French MT system was available from the HiTZ Center at the time of this study.

[14]https://elia.eus/itzultzailea

[15]https://www.batua.eus/

[16]https://www.euskadi.eus/itzuli/

	EU-ES	EU-FR
Llama 3.1 8B Instruct	0.066	0.158
Latxa 3.1 8B Instruct	0.270**	0.311***
Llama 3.1 70B Instruct	0.215*	0.254**
Latxa 3.1 70B Instruct	0.409***	0.352***
DeepSeek-V3.2-Exp	0.259**	0.275**
Claude 4 Sonnet	0.191*	0.166
GPT 5	**0.487*****	0.424***
SalamandraTA 7B Instruct	0.276**	0.354***
MADLAD-400-3B-MT	0.385***	0.195*
NLLB-200-3.3B	0.440***	**0.430*****
HiTZ MT eu-es	0.369***	–
Google Translate	0.390***	0.359***
Batua	0.320***	0.280**
Itzuli	0.284**	0.258**
Elia	0.246**	0.243**

Table 2: Pearson correlation coefficients (r) between Lanbide's labour statistics and model translations from WINOMT$_{eus}$ to Spanish and French. The highest correlation per language pair is highlighted in bold. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Occupation	Real female proportion (%)	GRAPE-M	
		ES	FR
housekeeper	96.5	23.4	24.6
tailor	92.7	12.4	12.5
receptionist	89.3	7.7	7.9
cashier	86.8	6.9	6.6
hairdresser	86.2	5.9	5.9

Table 3: Top five most feminized occupations in the real-world data showing the highest masculine bias scores (GRAPE-M) across all translation models.

Masculine overrepresentation. To complement the correlation analysis, we next examined the direction and magnitude of gender bias using the GRAPE metric. Results from the WINOMT$_{eus}$ experiment show that all models systematically favour masculine realizations of occupations, both in Spanish and French. The occupation most consistently overrepresented as masculine is *housekeeper*, despite being highly feminized in the Basque labour market (96.5% female). In Spanish, **Claude 4 Sonnet**, **Elia**, and **DeepSeek-V3.2-Exp** deviate the most from the expected gender distribution for this profession, whereas **Latxa 3.1 70B**, **NLLB-200-3.3B**, and **GPT 5** produce more faithful translations. Other highly feminized professions—such as *tailor*, *receptionist*, *cashier*, *hairdresser*, and *salesperson*—also exhibit strong masculine preferences, with similar GRAPE-M values in both target languages (see Tab. 3). Instances of feminine

overrepresentation are rare. Only ten cases display GRAPE-F values above 1.0, and most correspond to translation errors or lexical ambiguities. For example, in Spanish, **MADLAD-400-3B-MT** attains the highest GRAPE-F score (3.09) for *ertzain* (*Basque police officer*), often mistranslated as *la Ertzaintza*, referring to the institution rather than an individual. The only occupation that consistently appears predominantly in the feminine form across several models is *nurse*, with GRAPE-F values of 0.18 in Spanish and 0.16 in French. However, these scores remain close to zero, feminine overrepresentation being very weak compared to the masculine one observed earlier.

Limited gender effects on translation quality. Using FLORES+Gender, we evaluated whether translation quality differs across masculine and feminine sentence variants. We computed the average difference in translation quality between masculine and feminine subsets using the chrF++ and TER metrics and assessed the statistical significance of the observed differences through paired approximate randomization. At first glance, results in Tab. 4 suggest that, in general, models translate better when the referent is masculine in the Spanish-Basque direction. Nevertheless, most differences are small and not significant. The only exception is **Batua**, which exhibits significant differences in both metrics, consistently achieving better scores on masculine sentences. On the contrary, in the case of English as the source language, **NLLB-200-3.3B** shows a significant difference in the opposite direction in terms of chrF++, suggesting slightly better performance on feminine sentences.

Impact of annotated linguistic phenomena. To carry out a more detailed analysis, we used the annotations of the dataset to explore whether gender-related patterns observed persist under different conditions, such as the presence of multiple gendered entities, gendered proper names or the use of unmarked masculine forms. The values reported in Tab. 5 correspond to the difference between masculine–feminine performance when each factor is present compared to when it is not.

In our analyses, most gender-related differences were not statistically significant; however, several systems show significant or systematic tendencies under specific conditions. When sentences include multiple gendered entities, **HiTZ MT** tends to yield higher scores for masculine contexts in either language pair (e.g., *Es poco probable que el conductor* $_{\text{M}}$ *del vehículo que embistió al fotógrafo* $_{\text{M}}$ *sea procesado penalmente*, or *Liggins* $_{\text{M}}$ *followed in his* $_{\text{M}}$ *father* $_{\text{M}}$*'s footsteps and entered a career in medicine*). **Itzuli** also achieves significantly better (lower TER) performance in masculine multi-entity settings when translating from Spanish. Regarding proper names, **Elia** and **NLLB-200-3.3B** perform significantly better with feminine referents when translating from Spanish, as in *Después de eso, Kavita Paudwal* $_{\text{F}}$ *pasó al frente para cantar los bhajans*. Other systems show no clear advantage for either gender. In contrast, when translating from English, four models—**Latxa-70B**, **DeepSeek-V3.2-Exp**, **Batua** and **Itzuli**— show significant advantages for masculine names. Among the examined variables, the use of unmarked masculine (only applicable in Spanish) had the most consistent impact. Systems like **NLLB-200-3.3B**, **Itzuli** and **Elia** achieved better results for sentences containing unmarked masculine forms compared to their feminine counterparts (e.g., *Los parisinos* $_{\text{M}}$ *tienen reputación de egocéntricos, descorteses y engreídos*).

	ES-EU		EN-EU	
	chrF++	TER	chrF++	TER
Llama 3.1 8B Instruct	0.22	-21.07	0.00	-60.71
Latxa 3.1 8B Instruct	0.14	-0.50	-0.11	0.60
Llama 3.1 70B Instruct	0.20	3.01	-0.42	-0.44
Latxa 3.1 70B Instruct	0.14	0.11	-0.22	-0.44
DeepSeek-V3.2-Exp	-0.24	-0.29	-0.59	1.22
Claude Sonnet 4	-0.12	0.16	**-1.08***	-0.53
GPT 5	-0.36	0.50	-0.25	0.92
SalamandraTA 7B Instruct	0.20	-0.33	-0.03	-0.36
MADLAD-400-3B-MT	0.19	0.35	-0.13	-1.13
NLLB-200-3.3B	-0.49	-0.05	**-1.50***	-0.87
HiTZ MT	0.24	0.22	-0.14	1.27
Google Translate	0.26	**0.85***	0.01	0.77
Batua	**0.43***	**0.97***	-1.00	-0.82
Itzuli	0.22	0.63	-0.28	0.28
Elia	0.17	0.48	-0.29	0.45

Table 4: Average difference (Δ) between masculine and feminine contrastive sets from FLORES+Gender. Positive values (green) indicate higher performance on the masculine set, and negative values (orange) indicate higher performance on the feminine set. Statistically significant results are marked in boldface. $^{*}p < 0.05$.

7. Discussion

The results indicate that current MT systems and LLMs continue to reproduce gender asymmetries. In translation from Basque into gendered target languages, all evaluated systems show a systematic preference for masculine realizations of gender-neutral occupational terms. This behaviour is consistent with earlier reports of related work on other language pairs (Gete and Etchegoyhen, 2024; Mastromichalakis et al., 2025; Vanmassenhove et al., 2018; Schiebinger, 2014; Monti, 2020) and reflects the grammatical conventions of target languages where masculine forms function as the unmarked default, such as Spanish and French.

	ES-EU						EN-EU			
	ME		PN		UM		ME		PN	
	chrF++	TER	chrF++	TER	chrF++	TER	chrF++	TER	chrF++	TER
Llama 3.1 8B Instruct	0.44	-130.60	-0.28	-33.64	1.00	34.65	1.02	-65.17	2.06	0.61
Latxa 3.1 8B Instruct	-0.74	0.17	-0.33	-1.74	0.07	1.29	0.65	-0.75	0.43	0.75
Llama 3.1 70B Instruct	0.08	-28.39	-0.26	-31.92	0.07	-6.83	**1.57***	2.52	0.92	2.84
Latxa 3.1 70B Instruct	-0.28	0.31	0.08	0.43	-0.12	-0.24	1.79	2.22	**1.87***	2.73
DeepSeek-V3.2-Exp	0.80	2.03	0.21	1.52	-0.26	-0.81	0.00	1.32	1.08	**4.23****
Claude Sonnet 4	0.38	0.00	-0.46	-0.94	0.12	0.27	1.34	0.91	-1.14	-1.27
GPT 5	0.53	0.19	-0.36	-0.17	0.54	1.23	1.55	1.62	-0.44	-0.30
SalamandraTA 7B Instruct	-0.10	1.45	-0.55	-0.68	0.45	-0.55	-0.14	1.64	-1.68	-0.80
MADLAD-400-3B-MT	0.42	1.17	-0.62	-2.23	0.24	0.26	0.61	2.09	-1.12	-2.42
NLLB-200-3.3B	-0.07	0.36	**-1.44***	-1.09	**1.88****	**2.05****	0.59	0.00	-0.19	2.16
HiTZ MT	**0.84****	0.58	-0.34	-1.12	0.56	1.01	**2.27****	**4.53****	0.80	2.00
Google Translate	-0.31	0.93	0.36	0.66	0.30	0.47	1.59	1.58	0.85	2.22
Batua	0.00	-0.07	0.10	-0.52	-0.41	-0.08	1.27	2.99	0.58	**3.30***
Itzuli	0.50	**1.59***	-0.42	0.73	**1.05****	0.33	0.33	2.00	0.15	**2.38***
Elia	-0.21	-1.18	**-1.00***	**-2.43****	**1.19****	**1.89****	1.54	2.53	0.78	1.05

Table 5: Interaction effect (Δ) between gender and linguistic factors (ME = Multi-entity, PN = Proper names, UM = Unmarked masculine) across both Spanish-Basque (ES-EU) and English-Basque (EN-EU) translation directions. Green indicates higher performance on masculine sentences, orange indicates higher performance on feminine ones. Statistically significant results are marked in boldface. $^{**}p < 0.05$, $^{*}p < 0.1$

Although all models display a systematic preference for masculine realizations, the degree of bias varies across occupations. Professions that are predominantly feminine in real data (e.g., *housekeeper*, *tailor*, *receptionist*) are still often translated in the masculine—with a few exceptions, notably *nurse*—, yet the moderate correlations with Basque labour statistics indicate that models do capture some aspects of real-world gender distributions. In other words, while translation outputs reflect existing occupational asymmetries to a limited extent, they systematically exaggerate the masculine default due to its higher frequency and unmarked grammatical status in the training data.

The analysis of the translation *into* Basque using FLORES+Gender revealed a weaker and less systematic gender effect. For Spanish sources, most models produced slightly but rarely significant higher scores for masculine variants, particularly in sentences containing the generic masculine (e.g. *los investigadores*, 'the researchers') compared to their marked feminine counterparts (*las investigadoras*, 'the *female* researchers'). These results mirror the bias observed above, and suggest that the unmarked status and higher frequency of masculine forms in Spanish data may influence translation quality. In contrast, translations from English, a weakly gendered language, showed no consistent direction of bias: a few models achieved marginally better results for feminine sentences, while other performed slightly better with masculine proper names or multi-entity contexts.

8. Conclusions

This work introduces two high-quality resources for evaluating gender bias in translation involving Basque, a low-resource and genderless language. WinoMT$_{eus}$ adapts the WinoMT benchmark to Basque and provides a manually post-edited and culturally localised dataset that enables systematic assessment of how gender-neutral mentions of occupations are rendered in gendered target languages. Further, each occupation mention in WinoMT$_{eus}$ has been manually aligned with official Basque labour statistics, allowing researches to compare the gender distributions that emerge in model-generated translations against real-world employment data. This design makes it possible to quantify how translation systems reflect or diverge from actual societal gender patterns, an evaluation dimension rarely supported in previous resources.

FLORES+Gender, in turn, extends the FLORES+ benchmark with manually annotated gender-controlled contrastive pairs in Spanish and English. Each sentence has been validated for semantic equivalence and enriched with linguistic annotations (i.e., multiple entities, proper names, and unmarked masculine), creating a valuable testbed for analysing how grammatical gender in the source text affects translation quality into Basque.

Our experiments confirmed that current MT and LLM systems still default to masculine forms, even when the occupations are clearly associated with women. Regarding translation quality, results were less consistent: translations from Span-

ish, a strongly-gendered language, tended to be slightly better for masculine referent in some models, while no clear pattern emerged for English, a weakly-gendered language. Future work will leverage these resources to compare bias patterns across sociocultural contexts and to inform bias-aware training and evaluation practices involving the Basque language.

9. Acknowledgements

This work was supported by the HiTZ Chair of Artificial Intelligence and Language Technology (TSI100923-2023-1), funded by MTDFP, *Secretaría de Estado de Digitalización e Inteligencia Artificial*. Additional support was provided by the Research Project PID2024-157855OB-C32 (MOLVI), funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF), EU. It was also funded by the Basque Government (IKER-GAITU project) and the *Ministerio para la Transformación Digital y de la Función Pública* - Funded by EU – NextGenerationEU within the framework of the project *Desarrollo de Modelos ALIA*.

10. Ethics Statement and Limitations

Our work has several limitations that should be taken into account. First of all, the LLMs we evaluated show considerable variability, that is to say, small modifications in prompt design can lead to notable differences in the generated translations. Additionally, these models are typically used with relatively high values of temperature, which further increases randomness in their outputs. Fixing temperature at 0 allows to reproduce the experiments, but fails to reflect the variability present in real-world applications.

Regarding the newly created datasets, on the one hand, WINOMT$_{eus}$ focuses on a specific type of representational harm—misrepresentation—and covers a narrow domain, occupational representations, which captures only one aspect of gender bias in language technologies. Furthermore, this dataset is template-based, which is useful for controlled experimentation, but it does not represent the real use of language. FLORES+Gender, on the other hand, is based on real texts, but still presents some limitations. The size of both English and Spanish gendered sets is relatively small, and this may limit the reliability of the results. Apart from the size, this dataset also includes proper names to infer gender, which must be used carefully, given that relying on personal names without critical consideration can introduce or reinforce unintended biases. Translation quality was assessed through automatic metrics, which may not reflect the nuances of gender-related differences.

Another important limitation is that our approach treats gender as a binary variable (masculine/feminine). Although this choice reflects the grammatical systems of the analysed languages, it disregards non-binary and gender-neutral identities and provides an oversimplified view of gender in language.

Lastly, it should be noted that bias is a complex phenomenon that can manifest in many different ways depending on the task, domain or social context in which the models are used. Therefore, we cannot rely on a single NLP task, bias domain, metric or dataset to determine whether a model is biased. A high or low bias score in our datasets only should not be interpreted as evidence of generalized bias or the absence of it, but as an indication of how the model behaves in specific tasks and experimental conditions.

11. Bibliographical References

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015, Online. Association for Computational Linguistics.

Alessandra Teresa Cignarella, Anastasia Giachanou, and Els Lefever. 2025. A survey on stereotype detection in natural language processing. *ACM Computing Surveys*.

Ander Corral and Xabier Saralegi. 2022. Gender bias mitigation for NMT involving genderless languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 165–176, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marta R. Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale.

Kate Crawford. 2017. The trouble with bias. Keynote at NeurIPS.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872. ACM.

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.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models.

Harritxu Gete and Thierry Etchegoyhen. 2024. Does context help mitigate gender bias in neural machine translation? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14788–14794, Miami, Florida, USA. Association for Computational Linguistics.

Harritxu Gete, Thierry Etchegoyhen, Gorka Labaka, Ander Corral, Xabier Saralegi, Nora Aranberri, David Ponce, Igor Ellakuria Santos, and Maite Martin. 2025. Tando^{+}: corpus and baselines for document-level machine translation in Basque–Spanish and Basque–French. *Language Resources & Evaluation*.

Harritxu Gete, Thierry Etchegoyhen, David Ponce, Gorka Labaka, Nora Aranberri, Ander Corral, Xabier Saralegi, Igor Ellakuria, and Maite Martin. 2022. TANDO: A corpus for document-level machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3026–3037, Marseille, France. European Language Resources Association.

Javier Garcia Gilabert, Xixian Liao, Severino Da Dalt, Ella Bohman, Audrey Mash, Francesca De Luca Fornaciari, Irene Baucells, Joan Llop, Miguel Claramunt Argote, Carlos Escolano, and Maite Melero. 2025. From salamandra to salamandrata: Bsc submission for wmt25 general machine translation shared task.

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, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena

Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation.

Katsumi Ibaraki, Winston Wu, Lu Wang, and Rada Mihalcea. 2024. Analyzing occupational distribution representation in Japanese language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 959–973, Torino, Italia. ELRA and ICCL.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. Kobbq: Korean bias benchmark for question answering.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Johanna Monti. 2020. Gender issues in machine translation: An unsolved problem? In Luise von

Flotow and Hala Kamal, editors, *The Routledge Handbook of Translation, Feminism and Gender*, pages 457–468. Routledge, London and New York.

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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Abigail Oppong, Hellina Hailu Nigatu, and Chinasa T. Okolo. 2025. Examining the cultural encoding of gender bias in LLMs for low-resourced African languages. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 358–378, Vienna, Austria. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Spencer Rarrick, Ranjita Naik, Sundar Poudel, and Vishal Chowdhary. 2024. Gate x-e : A challenge set for gender-fair translations from weakly-gendered languages.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution.

Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, et al. 2025. Instructing large language models for low-resource languages: A systematic study for basque. *arXiv preprint arXiv:2506.07597*.

Abel Salinas, Parth Shah, Yuzhong Huang, Robert McCormack, and Fred Morstatter. 2023. The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA. Association for Computing Machinery.

Londa Schiebinger. 2014. Scientific research must take gender into account. *Nature*, 507:9–9.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ander Soraluze, Olatz Arregi, Xabier Arregi, Arantza Díaz de Ilarraza, Mijail Kabadjov, and Massimo Poesio. 2016. Coreference resolution for the Basque language with BART. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 67–73, San Diego, California. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational biases in Norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211, Seattle, Washington. Association for Computational Linguistics.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Pardis Sadat Zahraei and Ali Emami. 2025. Translate with care: Addressing gender bias, neutrality, and reasoning in large language model translations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 476–501, Vienna, Austria. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods.

Muitze Zulaika and Xabier Saralegi. 2025. BasqBBQ: A QA benchmark for assessing social biases in LLMs for Basque, a low-resource language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4753–4767, Abu Dhabi, UAE. Association for Computational Linguistics.