Diving into gender translation bias for the Portuguese language

Anonymous ACL submission

Abstract

Bias in Machine Translation models has become a significant concern. Despite extensive research in several language pairs, Portuguese remains under-explored. This study investigates gender bias in English-to-Portuguese translation. We conduct several experiments, extending an established dataset, and provide a comparative analysis of commercial Machine Translation systems, general-purpose LLMs, and non-commercial translation-specific models across various dimensions of gender bias. Additionally, we compare gender bias in Portuguese translation with that in other Romance languages (French, Spanish, and Italian). Finally, we explore whether sentiment influences gender bias in English-to-Portuguese translation.

1 Introduction

There has been little to no research on bias related to translation into Portuguese. Although several studies focus on Romance languages, the emphasis has been on French (Gonen and Webster, 2020; Stanovsky et al., 2019), Italian (Vanmassenhove, 2024a; Stanovsky et al., 2019), and Spanish (Gonen and Webster, 2020; Attanasio et al., 2023; Stanovsky et al., 2019). While findings for these languages are expected to be similar to Portuguese, biases manifest in distinct and unique ways for each language (Zhao et al., 2024). Therefore, studying these biases in the translation to Portuguese is a valuable and necessary task.

In this paper, we investigate gender bias in Portuguese translation, particularly the stereotypical associations between occupations and genders. For instance, words like "nurse" are frequently translated into feminine forms, while "doctor" is almost exclusively translated as masculine (Prates et al., 2020; Ghosh and Caliskan, 2023). On top of undermining the accuracy of the translated text, these biases can perpetuate and reinforce gender norms that contribute to the discrimination and marginalization of groups of people (Savoldi et al., 2021, 2024).

To evaluate the presence of gender biases in translation, we employ and extend WinoMT (a benchmark for co-reference resolution of occupational referents), to compare state-of-the-art commercial Machine Translation (MT) systems, general-purpose Large Language Models (LLMs), and non-comercial translation-specific models. Our contributions focus on addressing gender bias in English-Portuguese translations by answering the following questions:

- Which system performs best, considering several bias metrics, when translating single sentences? Do these systems tend to over rely on stereotypes to make gender predictions? In addition to evaluating translation bias, we assess overall translation quality and also conduct human evaluations to ensure the reliability of these results.

- How do these systems behave when translating two-sentence contexts, where pronoun referents appear in a different sentence? And if an intermediate sentence separates the two? Existing datasets only allow for the study of translation bias within single sentences (Menezes et al., 2023). To advance inter-sentence bias translation evaluation, we propose two extensions of an existing dataset: one enabling the analysis of bias in a two-sentence context (inter-2) and another incorporating an intermediate sentence (inter-3).

- How does translation bias into Portuguese compare with other Romance languages (French, Spanish, and Italian)? Here, we update previous gender translation bias results' for these languages (Stanovsky et al., 2019), and evaluate current state-of-the-art models in

these languages, while comparing them with Portuguese.

- Can we say that the sentiment of sentences influences gendered translations in Portuguese? Prior research (Stanovsky et al., 2019; Prates et al., 2020) demonstrates that certain adjectives, such as "shy" or "proud", can influence gender outcomes in translations. Cho et al. (2019) and Cho et al. (2021) also explore this effect, but with a focus on the sentiment conveyed by the adjectives. Building on this idea, we investigate whether the overall sentiment of a sentence influences gender in Portuguese translations.

Our results show that commercial MT systems still lead in producing unbiased translations. However, we emphasize that traditional quality metrics fail to capture significant biases present in these systems. Our study also reveals inter-sentence bias as a striking weakness of current models. Furthermore, while our findings suggest that commercial MT systems have improved over the past few years, translations from French, Spanish, and Italian still exhibit far worse results in terms of gender bias compared to Portuguese translations. Finally, we found no strong evidence that sentiment significantly influences the translated gender of sentences.

2 Related Work

Stanovsky et al. (2019) present the first large-scale multilingual evaluation of gender bias in MT and introduce the WinoMT dataset. Their methodology leverages co-reference resolution datasets to assess whether MT systems can accurately translate gendered roles without defaulting to stereotypical biases, and measures how each system's performance varies based on gender and stereotypical role assignments. Their experiments conclude that all the tested MT systems are gender biased. Although this study was from 2019, it was pivotal in providing tools and a concrete evaluation method to study gender bias in MT. Their methodology continues to be used in current research to evaluate the presence of bias (Basta et al., 2020; Attanasio et al., 2023; Stafanovičs et al., 2020; Saunders and Byrne, 2020) and the effectiveness of gender bias mitigating strategies (Saunders et al., 2020; Saunders and Byrne, 2020; Stafanovičs et al., 2020). Our work also builds on this foundation.

Prates et al. (2020) and Cho et al. (2019) employ similar methodologies to evaluate gender bias in the translation of neutral pronouns into English, utilizing occupations and sentiment words as contextual information. Prates et al. (2020) use sets of sentences structured as "He/She is <occupation>" across 12 genderless languages to measure the frequency of female, male, and neutral pronouns in the translations for each occupation. Upon comparing their results with data from the U.S. Bureau of Labor Statistics, the study concludes that Google Translate's preference for male defaults does not correlate with unequal representation of female and male workers in those occupations. Cho et al. (2019) extend this approach to Korean-English translations, reaching a similar conclusion regarding a masculine bias in translations.

Also taking advantage of simple template sentences, Cho et al. (2021) explore how occupations and sentiment words are translated between languages with different gender systems. They employ phrases such as "One thing about the man/woman, he/she is <occupation/sentiment word>". To the best of our knowledge, this is the only study to include Portuguese in the evaluation of different translation systems. While their work provides valuable insights about the persistence of gender bias, their analysis of Portuguese remains limited. Our work aims to expand on this by offering a more in-depth investigation of gender bias in Portuguese translations, focusing on a broader set of linguistic structures and models.

Trainotti Rabonato et al. (2024) also focus on the Portuguese language, but their work targets bias mitigation strategies, namely fine-tuning. To evaluate the efficiency of this strategy in the chosen model, MarianMt, they adopt the same foundational methodology as our work, established by Stanovsky et al. (2019). However, their evaluation only assesses the model being improved, without investigating how it compares to other publicly available systems. In contrast, our contribution lies in conducting a comprehensive evaluation of the current state of gender translation bias for Portuguese.

With the wide adoption of LLMs, new research has been exploring the gender bias present in these models's outputs for different languages. Ghosh and Caliskan (2023) and Vanmassenhove (2024b) evaluate LLMs, specifically GPT models, in the concrete task of translation. Ghosh and Caliskan (2023) focus on the translation of low-resource languages that exclusively use gender-neutral pronouns, such as Bengali. Their investigation exposes

gender biases in the portrayal of occupations (man = doctor, woman = nurse) and actions (woman = cook, man = go to work). It also reveals the inability to translate the English gender-neutral pronoun *they* into equivalent gender-neutral pronouns in these languages. Similarly, Vanmassenhove (2024b)'s findings reinforce ChatGPT's inability to handle gender in a systematic manner. When prompted to provide gender alternatives for translations from Italian to English, the model often falls short and may even exhibit additional biases.

In this paper, we advance the state of the art by presenting a detailed evaluation of current widely used models regarding gender bias in translations from English to Portuguese and compare the results with those for other Romance languages.

3 Methodology

We followed the methodology outlined for the aforementioned WinoMT test suite (Stanovsky et al., 2019), which consists of the following main steps that were reproduced for every model: a) **Translation**, b) **Alignment**, and c) **Gender Extraction**

3.1 Translation

The first step is to translate all sentences in the dataset into Portuguese using a target model. The models we tested are:

1. **Commercial MT systems**: Google Translate[1], Amazon Translate[2], Microsoft Translate[3], DeepL[4];

2. **General-purpose LLMs**: GPT-4o[5], DeepSeek-V3[6] (Liu et al., 2024), Llama3.2[7], Llama3.2 Instruct[8] (Dubey et al., 2024), Tower Base[9], Tower Instruct[10] (Alves et al., 2024), EuroLLM[11] (Martins et al., 2024);

3. **Translation-specific non-comercial models**: NLLB-200[12] (Costa-jussà et al., 2022), M2M[13] (Fan et al., 2021), OPUS-MT[14] (Tiedemann and Thottingal, 2020).

For the commercial systems, we utilized the APIs provided by each service. For the remaining models, we accessed them through the Hugging Face interface.

When the models allowed us to specify between European Portuguese and Brazilian Portuguese, we always opted for European Portuguese. However, for some models, this option was unavailable and translations would occasionally default to Brazilian Portuguese. We also encountered an issue with the Google Translate API where even if European Portuguese was specified, the API returned translations in Brazilian Portuguese. In those cases, we decided to proceed with the Brazilian Portuguese translations, assuming that any gender bias would likely be comparable between the two varieties.

Regarding prompt-based LLMs such as GPT, Llama, and Tower Instruct, we employed a zero-shot approach. The basic structure of our prompts was as follows, with slight variations to suit each model's expected response format:

```
Translate the following text from English
into Portuguese.
English: {sentence}
Portuguese:
```

Table 1 presents an example sentence and the translations produced by different MT systems. We can observe the various approaches each system uses to handle both the entity with explicit gender reference and the ambiguous entity. In our work, we evaluate only the entity whose gender is unambiguous given the context (in Table 1, the librarian). For the other entity in the sentence, determining the best way to translate gender-ambiguous entities is a subject of ongoing discussion, and each translation system handles such cases differently. Although biases may influence these decisions, addressing them is beyond the scope of our evaluation, and our focus remains on clear-cut biases that lead to unequivocally incorrect sentences.

[1] https://translate.google.com
[2] https://aws.amazon.com/translate
[3] https://www.bing.com/translator
[4] https://www.deepl.com/en/translator
[5] https://chatgpt.com
[6] https://www.deepseek.com/
[7] https://huggingface.co/meta-llama/Llama-3.2-3B
[8] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[9] https://huggingface.co/Unbabel/TowerBase-7B-v0.1
[10] https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2
[11] https://huggingface.co/utter-project/EuroLLM-9B
[12] https://huggingface.co/facebook/nllb-200-3.3B
[13] https://huggingface.co/facebook/m2m100_1.2B
[14] https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-en-pt

Source Sentence:
The librarian was unable to find the book for **the developer** and instead offered **her** a magazine. (developer, female)

System	Predicted translation	Phenomenon
Google Translate	**A bibliotecária** não conseguiu encontrar o livro para **a desenvolvedora** e, em vez disso, ofereceu a **ela** uma revista.	Translate all entities to same gender (female).
DeepL	**O bibliotecário** não conseguiu encontrar o livro para **a promotora** e ofereceu-**lhe** uma revista.	Correct gender for the target entity and defaults to male for the ambiguous entity (bibliotecário). Note: Incorrect translation of developer as "promotora".
M2M	**O bibliotecário** não conseguiu encontrar o livro para **o desenvolvedor** e, em vez disso, ofereceu-**lhe** uma revista.	**Biased** translation, defaults both entities to masculine forms, ignoring gender context in the source sentence.
GPT	**O bibliotecário** não conseguiu encontrar o livro para **o desenvolvedor** e, em vez disso, ofereceu a **ela** uma revista.	**Biased** translation, defaults both entities to masculine but uses a feminine pronoun ("ela") to address the developer.
Llama-Instruct	**A bibliotecária** não conseguiu encontrar o livro para **o desenvolvedor** e, em vez disso, ofereceu-**lhe** uma revista.	**Biased** translation, likely influenced by stereotypical roles.

Table 1: Examples of the different systems' performance on a sentence from the WinoMT corpus. Words in **blue**, **orange** and red indicate male, female and neutral, respectively.

3.2 Alignment

The next step involves aligning the source sentences and their Portuguese translations. This step matches occupational nouns (e.g. *the lawyer*) in the original sentences with the corresponding ones in the translated text (e.g. *o advogado*). Given the limited resources available for English-Portuguese text alignment, we opted to use *fast-align* (Dyer et al., 2013), which is known for its efficiency.

3.3 Gender Extraction

Finally, a morphological analyzer is employed to extract the gender of entities in the target side. This allows us to compare the gender of the translated entity against the correct gender assigned by the annotations.

We tested the default Portuguese models for SpaCy 2.2[15] (Honnibal et al., 2020), Stanza 1.9 (Qi et al., 2020) and UDPipe 2.5 (Straka et al., 2016) in the Bosque dataset (Rademaker et al., 2017) from the Universal Dependencies framework (de Marneffe et al., 2021). Stanza performed the best in terms of overall gender accuracy, but due to its considerably longer runtime compared to SpaCy, which achieved similar results, we opted to use SpaCy in our experiments.

None of the analyzers demonstrated a significant difference in performance between feminine and masculine words, which is important to minimize the risk of inadvertently introducing new bias into our research.

4 Experimental Setup

4.1 Datasets

4.1.1 Single-sentence dataset

As previously stated, in this study we rely on the WinoMT dataset, introduced by Stanovsky et al. (2019), which results from two English benchmarks for co-reference resolution: WinoBias (Zhao et al., 2018) and WinoGender (Rudinger et al., 2018). WinoMT is commonly used to assess model's ability to resolve pronoun references, using occupations as contextual information. It consists of 3,888 sentences featuring two human entities, defined by their occupations, along with a pronoun referring to one of them. It is equally balanced between female and male genders, as well as pro-stereotypical and anti-stereotypical role assignments. Each sentence is annotated with the entity

[15] Although this is not the most recent version of SpaCy, SpaCy v2 obtained better gender accuracy than SpaCy v3 in our experiments.

to be tested and its corresponding gender label.

Example:

Pro-stereotypical: *The **developer** argued with the **designer** because he did not like the design.*

Anti-stereotypical: *The **developer** argued with the **designer** because she did not like the design.*

4.1.2 Inter-sentence datasets

We also explored inter-sentence contexts, where the gender clues are provided in a separate sentence.

We manually selected a set of 500 sentences from the original dataset that could be easily split into two parts without loosing meaning or creating ambiguity. In most cases, this division was done by splitting the sentences at the word "because".

Example:

Original: *The developer visited the hairdresser because she needed to cut her hair.*

Divided: *The developer visited the hairdresser. She needed to cut her hair.*

Example of a discarded sentence, due to ambiguity:

Original : *The chief gave the housekeeper a tip because she was satisfied.*

Divided: *The chief gave the housekeeper a tip. She was satisfied.*

Additionally, we made sure that the 500 sentences were evenly balanced between male and female references, as well as between pro-stereotypical and anti-stereotypical role assignments. We call this dataset **inter-2**.

We also created an additional dataset by adding a new neutral sentence in between the other two. We experimented with two different options: "It made sense" and "The weather was cold.".

Example:

Version 1: *The CEO bought the accountant a car. It made sense. He needed one.*

Version 2: *The CEO bought the accountant a car. The weather was cold. He needed one.*

We call this dataset **inter-3**.

4.2 Metrics

To measure the impact of gender bias in translations, we used the same metric implementation as the WinoMT test suite, with the addition of the masculine-to-female (M:F) ratio (Saunders and Byrne, 2020). The metrics are as follows:

- **Accuracy**: the percentage of instances the translation has the correct gender.
- ΔG: difference in performance (F_1 score) between male and female translations.
- ΔS: difference in performance (Accuracy) between pro-stereotypical and anti-stereotypical role assignments.
- **M:F ratio**: ratio of male and female predictions.

While accuracy provides a general sense of system performance and existing bias, the other metrics allow for a more fine grained analysis of where these gender biases reside.

Smaller values of ΔG and ΔS are indicative of less gender bias. A high ΔG suggests that the system performs better for one gender over the other. Meanwhile, a high ΔS indicates that the system performs poorly when handling anti-stereotypical roles, pointing to an over-reliance on stereotypes.

The M:F ratio should be as close to 1 as possible, as the WinoMT dataset is equally balanced between masculine and feminine genders. This metric correlates with ΔG, but also allows for a more nuanced analysis of the other metrics. If M:F ratio is skewed (either too high or too low), it reduces the relevance of ΔS, as the imbalance would suggest that the system is more heavily influenced by one gender.

5 Results and Discussion

5.1 Single Sentence Results

5.1.1 Automatic Evaluation

Our main findings are presented in Table 2. As in all tables of this paper, the best results for each column are shown in bold, while the worst results are shown underlined.

In terms of gender accuracy, commercial MT systems outperform all other models. Only NLLB and Tower Instruct can achieve comparable results.

Balanced gender performance is a notable strength of commercial MT systems, with ΔG scores close to zero. Interestingly, both Amazon Translate and Google Translate perform slightly better for feminine entities than for masculine ones. Tower Instruct stands out among the remaining MT systems with a relatively low ΔG, followed by NLLB. On the other end of the spectrum, DeepSeek is the worst-performing model, followed by OPUS-MT and M2M.

Stereotypical bias, highlighted by ΔS, remains a widespread issue. All systems perform significantly better on stereotypical sentences, which

Model	Acc	$\Delta G\uparrow$	$\Delta S\uparrow$	M:F$\downarrow$
Google Translate	78.8	**0.4***	17.7	1.46
Amazon Translate	**79.4**	0.8*	**9.8**	**1.39**
Microsoft Translate	73.7	1.1	18	1.48
DeepL	75.4	0.9	21.7	1.47
GPT-4o	59.8	9.8	31.8	2.10
DeepSeek-V3	47.3	35.2	25.2	5.13
Llama3.2 3B	57.9	14.3	30.3	2.59
Llama3.2 3B Instruct	56.1	20.9	25.2	3.38
TowerBase 7B	61.6	13.9	24.6	2.78
TowerInstruct 7B	71.9	1.5	24.9	1.43
EuroLLM 9B	64.4	18.8	11.9	2.76
OPUS-MT	54.8	27.4	21.6	4.57
NLLB200 3.3B	77.2	4	22.7	1.83
M2M100 1.2B	57.5	22	23.7	3.97

Table 2: Performance of various translation models on the WinoMT dataset. Values with * indicate that the imbalance is favoring female instances.

points to an over-reliance on gender stereotypes. Commercial systems once again exhibit the best results, but still show significant weaknesses in this regard.

The M:F ratio also reveals that all systems, to some extent, default to male terms. This is not surprisingly, as in Portuguese masculine forms are often used as the neutral or default form. Commercial MT systems again perform best with this metric, with Tower Instruct being the only LLM that not only approaches Amazon Translate but even outperforms other commercial MT models. DeepSeek, however, stands out for its poor performance in this metric, consistent with its similarly weak performance in ΔG. This suggests a very strong tendency to default to male forms, regardless of context or gender clues.

5.1.2 Human Evaluation

We conducted a human validation to estimate the accuracy of our automatic gender bias evaluation. We randomly sampled 60 sentences, selecting translations from different translation models, and had each sentence annotated by three native Portuguese speakers. The annotators were tasked with identifying the gender of an entity for each translations.

We then compared the human annotations with the automatic annotations. The agreement between human and automatic annotations was always above 89%. Additionally, the inter-annotator agreement among humans annotators was 91%. An example of discrepancy was the term *um aluno* ("a student") in the sentence *O educador estava se reunindo com um aluno para discutir suas habilidades de escrita.* ("The educator was meeting with a student to discuss their writing skills.") that was considered masculine by two annotators, while other classified it as neutral. This reflects the common practice in Portuguese of using masculine forms as generic or gender-neutral. Another source of confusion involved the translation of certain occupational terms, which were either incorrect or unfamiliar. For instance, "cashier" was translated as *o caixa* (masculine) or *a caixa* (feminine), and the term "janitor" was sometimes translated as *garçonete*. Since these terms are rarely used in Portugal, some annotators were uncertain about how to classify them.

5.1.3 Translation Quality

We evaluated translation quality using the COMET-Kiwi metric (Rei et al., 2022). Results are presented in Table 3.

Model	COMET Score(%)
Google Translate	**84.9**
Amazon Translate	83.3
Microsoft Translate	84.1
DeepL	84.3
GPT-4o	83.8
DeepSeek-V3	84.4
Llama-3.2 3B	81.4
Llama-3.2 3B Instruct	82.2
TowerBase 7B	84.5
TowerInstruct 7B	84.6
EuroLLM 9B	84.5
OPUS-MT	84.1
NLLB200 3.3B	84.4
M2M100 1.2B	83.2

Table 3: COMET scores for various translation models on the WinoMT dataset.

Most models achieve comparable results, with no model significantly outperforming the others. Several conclusions can be drawn from this observation. First, systems with different bias scores perform similarly in terms of translation quality, which underscores that it is possible to achieve fairer translations while maintaining high quality outputs. Second, as highlighted earlier in this work, even state of the art translation quality metrics alone are insufficient to capture biases.

5.2 Inter-Sentence Results

Considering the inter-sentence gender bias, Table 4 presents the accuracy of each model on the 500 original sentences, the inter-2 (I2) and inter-3 (I3) datasets. For the inter-3 dataset, the two different versions, "It made sense" and "The weather was cold.", produced similar results, therefore, we present only the results for the first one.

The full results for all remaining metrics are provided in Appendix A.

Model	Original	I2	I3
Google Translate	78.2	**75.2**	**71.2**
Amazon Translate	**85.0**	47.0	47.4
Microsoft Translate	78.6	46.4	46.4
DeepL	78.4	65.0	45.0
GPT-4o	47.4	47.8	51.8
DeepSeek-V3	47.4	46.8	47.0
Llama-3.2 3B	55.0	53.2	49.6
Llama-3.2 3B Instruct	55.4	60.2	54.0
TowerBase 7B	61.2	66.4	56.2
TowerInstruct 7B	74.2	73.2	67.2
EuroLLM 9B	70.6	71.8	67.0
OPUS-MT	56.8	55.4	54.4
NLLB200 3.3B	75.2	68.8	64.4
M2M100 1.2B	62.8	49.8	47.2

Table 4: Accuracy of all models on a portion of the original WinoMT dataset versus inter-2 and inter-3.

Most translation systems exhibit a high drop in accuracy when faced with two separate sentences (I2). Commercial MT systems are the most affected by this modification. Amazon Translate and Microsoft Translate, two of the best-performing models on the original sentences, show the most significant declines in accuracy. In contrast, some models perform slightly better with this modification, particularly general-purpose LLMs. This highlights their ability to retain and utilize contextual information spread across sentences.

Adding a sentence between the two others (I3) led to a decrease in accuracy for most models. Some models remained unaffected by this additional sentence, as they had already experienced a significant drop with the previous change. Among the models that maintained good performance with two sentences, DeepL showed the most noticeable decline with the addition of a third sentence.

Among the best performing systems on the original sentences, only Google Translate, EuroLLM, and Tower Instruct maintain high accuracy in the two new sets of sentences, indicating robustness in capturing inter-sentence context. Based on these results, Google Translate is the most effective model at handling cases where contextual information appears in a separate sentence.

5.3 Portuguese vs. other Romance Languages

We started by examine the original study that introduces the WinoMT test suite, conducted in 2019. After replicating the experiments with updated translations, our findings indicate an improvement over the results reported there for Romance Languages, shown on Table 5. This suggests that translation systems have made substantial effort in recent years to generate fairer translations.

Then, we conducted a comparative analysis across those Romance languages using all the previously evaluated translation systems[16]. Our evaluation included Spanish, French, and Italian. Results are shown on Table 6. Interestingly, the English-Portuguese language pair yielded significantly better results than the other tested languages, across all metrics.

5.4 Sentiment Analysis

The sentences in the WinoMT dataset often convey strong sentiments (e.g.,*The developer argued with the designer because she did not like the design.* [NEGATIVE], *The mover said thank you to the housekeeper because she is grateful.* [POSITIVE]). In this section, we analyze how this aspect influences the translation of the entities in the sentence.

To perform sentiment analysis, we utilized Hugging Face's sentiment analysis pipeline with the default DistilBERT model[17]. We then measured the correlation between sentiment (positive/negative) and the gender of translations (masculine/feminine), using Pearson's correlation coefficient. The results of this analysis are presented in Table 7.

Overall, we did not find strong evidence that sentiment plays a significant role in the translated gender of the sentences. Even in cases with statistically significant results (e.g., GPT, Google Translate, Llama, Tower Base), the correlations remain weak. This suggests that gender (male vs. female) is not meaningfully associated with sentiment polarity (positive vs. negative) in the evaluated translation models.

6 Conclusions

In this study, we explored the presence of gender bias in English-to-Portuguese translation. We found that that all the systems tested exhibit prevalent biases. While commercial MT systems outperform others across many metrics, they remain far from perfect, continuing to rely heavily on stereotypes. This issue persists even in cases where sufficient context is provided to accurately identify the gender of an entity.

[16]DeepL was excluded from this evaluation due to the free character limit of its API

[17]https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

	Google Translate			Microsoft Translate			Amazon Translate		
	Acc	$\Delta G\downarrow$	$\Delta S\downarrow$	**Acc**	$\Delta G\downarrow$	$\Delta S\downarrow$	**Acc**	$\Delta G\downarrow$	$\Delta S\downarrow$
ES	53.1	23.4	21.3	47.3	36.8	23.2	59.4	15.4	22.3
FR	63.6	6.4	26.7	44.7	36.4	29.7	55.2	17.7	24.9
IT	39.6	32.9	21.5	39.8	39.8	17.0	42.4	27.8	18.5

Table 5: Results for Spanish, French and Italian presented in the 2019 study by Stanovsky et al.

	Google Translate				Amazon Translate				Microsoft Translate			
	Acc	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$	**Acc**	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$	**Acc**	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$
PT	**77.8**	**0.4***	**17.7**	**1.46**	**79.4**	**0.8***	**9.8**	1.39	**73.7**	1.1	**18**	1.48
ES	64.8	6.7	28.4	2.13	73.8	1.8	18.8	**1.22**	71.9	**0.5***	20.7	**1.37**
FR	62.0	8.0	23.8	2.29	67.5	1.5	22.2	1.86	65.8	2.5	19.5	1.85
IT	50.5	6.6	25.3	2.04	54.0	7.9	19.5	2.19	48.6	13.8	23.6	2.63

	GPT-4o				DeepSeek-V3			
	Acc	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$	**Acc**	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$
PT	**59.8**	**9.8**	31.8	**2.10**	47.3	35.2	25.2	5.13
ES	56.5	11.8	33.9	2.39	**68.2**	3.7	25.7	1.84
FR	57.9	11.1	**19.3**	2.53	62.4	**0.6**	24.6	**1.55**
IT	43.7	17.9	19.8	2.94	38.5	30.4	**20**	5.13

	Llama-3.2 3B				Llama-3.2 3B Instruct			
	Acc	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$	**Acc**	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$
PT	**57.9**	**14.3**	30.3	**2.59**	**56.1**	20.9	25.2	3.38
ES	53.3	18	**18.4**	2.96	53.9	24.2	**19.5**	4.25
FR	52.2	17.1	20.5	2.86	54.1	**16.3**	27.6	**3.13**
IT	46.3	17.7	23	2.92	45.5	24.9	21.4	4.19

	TowerBase 7B				TowerInstruct 7B				EuroLLM 9B			
	Acc	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$	**Acc**	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$	**Acc**	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$
PT	**61.6**	**13.9**	24.6	**2.78**	**71.9**	**1.5**	24.9	1.43	64.4	18.8	**11.9**	2.76
ES	54.6	21.9	22.4	3.81	68.3	1.7	**20.7**	**1.42**	**64.6**	**7.9**	16.6	**2.22**
FR	51.7	23.4	**20.3**	4.06	64.4	2.8	23.4	1.60	58.0	12.7	18.4	2.87
IT	44.4	20.6	26.2	3.65	54.3	3.7	22.5	1.65	49.7	14.5	23.9	3.11

	OPUS-MT				NLLB200 3.3B				M2M100 1.2B			
	Acc	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$	**Acc**	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$	**Acc**	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$
PT	**54.8**	27.4	21.6	4.57	**77.2**	**4**	22.7	**1.83**	**57.5**	22	23.7	3.97
ES	54.4	20.9	24.6	3.69	68.5	4.1	28.4	1.88	**57.5**	**18.2**	20.7	**3.59**
FR	52.0	**20**	22.2	**3.52**	63.2	5.9	28.3	2.14	51.8	22.7	20.6	4.05
IT	40.5	28.3	**19.9**	4.72	54.8	6.1	**20.6**	1.96	42.9	26.1	**19**	4.39

Table 6: Performance of the translation models on the WinoMT dataset across different languages.

Model	**Correlation**	**p-value**
Google Translate	-0.053	0.001
Amazon Translate	-0.013	0.602
Microsoft Translate	0.005	0.843
DeepL	-0.016	0.534
GPT-4o	**-0.076**	**0.000**
DeepSeek-V3	-0.002	0.879
Llama-3.2 3B	-0.041	0.012
Llama-3.2 3B Instruct	-0.002	0.920
TowerBase 7B	-0.042	0.005
TowerInstruct 7B	-0.004	0.823
EuroLLM 9B	-0.015	0.358
OPUS-MT	-0.018	0.260
NLLB200 3.3B	0.002	0.899
M2M100 1.2B	-0.028	0.113

Table 7: Pearson Correlation and p-values for different models.

Additionally, we examined gender bias in an inter-sentence context. By dividing each sentence into two, we test the robustness of translation models to utilize contextual information about gender spread across sentences. Our findings indicated that very few models were able to maintain high performance when faced with this challenge. Performance decreased further when a neutral sentence was added between the two previous sentences.

We also compared the results for Portuguese with those of other Romance languages, revealing some surprising findings. Portuguese shows better performance than some other Romance languages. Furthermore, this experiment allowed us to revisit previously reported results for these languages, and conclude that there has been noticeable improvement in translation systems over time.

Finally, we concluded that there was no strong relation between the sentences' sentiment and the translated gender.

7 Limitations

In this work, our analysis was restricted to bias related to occupational stereotypes, which is only one of many manifestations of gender bias. Additionally, this study only considers binary gender forms (female and male), overlooking non-binary and gender-neutral representations. Given the complexity of this issue, including gender-neutral translation into our study would present significant challenges. Nevertheless, we acknowledge the importance of this endeavor and leave it open for future exploration. Moreover, the structured nature of the sentences in our dataset, designed to isolate occupational stereotypes, may lead models to learn these specific patterns without addressing broader issues of bias.

Also, due to hardware limitations, we were unable to test models with more than 10 billion parameters, which meant the largest versions of some models were excluded from the research.

It is also important to note that we cannot definitively determine whether the translation models have previously been exposed to the WinoMT dataset during training or testing. This would undermine the reliability of our findings in measuring the true extent of gender bias in these models, as their performance may not accurately represent their real-world behavior when confronted with new, unseen data.

Finally, the results presented in this work are tied to the specific versions of the models used during this study, some of which can be updated with at any time. As we have seen, some commercial MT systems have evolved since the first study to use this evaluation method, and will probably continue to do so. The use of LLMs also adds another layer of complexity as reproducibility with these models is a significant concern. The model's response to the same query may vary. Prompt design also plays a crucial role in determining the outputs of LLMs as variations in input prompts could produce different results, impacting the study's findings.

8 Ethics Statement

We acknowledge the sensitive nature of gender bias and took care to present our findings in a responsible manner. This study relied upon the WinoMT test suite, a publicly available resource that, to the best of our knowledge, was collected in accordance with all ethical standards. Similar to WinoMT, our expansion of the dataset will be made publicly available under the MIT license to facilitate reproducibility and further research in this area. For the human evaluation component, all participants were fully informed of the purpose of the research and participated voluntarily with full consent. No financial compensation was provided for participation. Additionally, ChatGPT was used as an editing tool to improve the clarity and coherence of the text in this paper.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pages 99–102, Seattle, USA. Association for Computational Linguistics.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Won Ik Cho, Jiwon Kim, Jaeyeong Yang, and Nam Soo Kim. 2021. Towards cross-lingual generalization of translation gender bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 449–457.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 644–648.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 901–912.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Miguel Menezes, M. Amin Farajian, Helena Moniz, and João Varelas Graça. 2023. A context-aware annotation framework for customer support live chat machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 286–297, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal dependencies for portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy.

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.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

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.

Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024. What the harm? quantifying the tangible impact of gender bias in machine translation with a human-centered study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076, Miami, Florida, USA. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

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.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ricardo Trainotti Rabonato, Evangelos Milios, and Lilian Berton. 2024. Gender-neutral english to portuguese machine translator: Promoting inclusive language. In *Brazilian Conference on Intelligent Systems*, pages 180–195. Springer.

Eva Vanmassenhove. 2024a. 9 gender bias in machine translation and the era of large language models. *Gendered Technology in Translation and Interpreting: Centering Rights in the Development of Language Technology*, page 225.

Eva Vanmassenhove. 2024b. Gender bias in machine translation and the era of large language models. *arXiv preprint arXiv:2401.10016*.

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.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages. *arXiv preprint arXiv:2403.00277*.

A Appendix

	Accuracy			$\Delta G\downarrow$			$\Delta S\downarrow$			M:F↓		
	Orig.	**I2**	**I3**	**Orig.**	**I2**	**I3**	**Orig.**	**I2**	**I3**	**Orig.**	**I2**	**I3**
Google Translate	78.0	75.2	71.2	0.1	1.7	4.6	22.1	23.1	22.6	1.43	1.57	1.77
Microsoft Translate	78.6	46.4	46.4	0.5	35.8	35.8	9.6	23.6	24.1	1.30	4.70	4.63
Amazon Translate	85.0	47.0	47.4	2.5	41.2	41.5	1.0	24.0	23.1	1.19	6.66	6.90
DeepL	78.4	65.0	45.0	0.1	8.5	36	16.8	31.3	28.8	1.38	2.09	4.89
GPT-4o	47.4	47.8	51.8	26.6	31	23.9	35.6	34.6	45.2	3.33	4.18	3.43
DeepSeek	47.4	46.8	47.0	36.8	36.2	38.1	29.3	30.8	29.3	5.24	5.07	5.65
Llama-3.2 3B	55.0	53.2	49.6	18.0	19.4	25.8	34.7	33.2	34.7	2.87	2.84	3.50
Llama-3.2 3B Instruct	55.4	60.2	54.0	20.4	15.7	22.5	31.7	21.2	27.4	3.10	2.80	3.41
TowerBase 7B	61.2	66.4	56.2	14.3	8.8	20.8	25.5	25.5	29.8	2.68	2.18	3.17
TowerInstruct 7B	74.2	73.2	67.2	0.3	0.1	3.3	26.9	29.4	28.4	1.23	1.21	1.31
EuroLLM 9B	70.6	71.8	67.0	6.9	5.1	9.8	17.8	12.9	18.7	2.28	1.93	2.46
OPUS-MT	56.8	55.4	54.4	25.6	25.8	27.9	23	25	20.2	4.47	4.38	4.67
NLLB200 3.3B	75.2	68.8	64.4	4.5	6.4	10.8	23.1	24.5	29.9	1.74	2.10	2.47
M2M100 1.2B	62.8	49.8	47.2	15	30.7	32.9	29.8	25.0	29.3	2.98	4.86	5.06

Table 8: Results for all metric on the datasets inter-2 and inter-3.