Diving into Gender Translation Bias for the Portuguese Language

A Comparative Analysis of Commercial Machine Translation Systems, General-Purpose LLMs, and Non-Commercial Translation-Specific Models

Sofia Seabra Bonifácio

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur
Prof. Helena Gorete Silva Moniz

Examination Committee

Chairperson: Prof. Daniel Jorge Viegas Gonçalves
Supervisor: Prof. Maria Luísa Torres Ribeiro Marques da Silva Coheur
Member of the Committee: Dr. Ana Catarina dos Santos Farinha

June 2025

This work was created using LaTeX typesetting language
in the Overleaf environment (www.overleaf.com).

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

Em primeiro lugar, gostaria de agradecer às minhas supervisoras, Luísa e Helena, por me terem dado a oportunidade de explorar este tema, pelo qual tenho tanto carinho. Obrigada por todo o vosso apoio e encorajamento, que me ajudaram a nunca perder a motivação, e pelo vosso conhecimento e interesse na área, que me permitiram aprender tanto. Foi um prazer trabalhar ao vosso lado.

Obrigada também aos meus pais, irmãos e restantes familiares, pelo apoio e paciência ao longo dos últimos meses. Obrigada por sempre demonstrarem tanto interesse pela minha tese e por estarem lá mesmo nos dias mais stressantes.

Obrigada a todos os amigos que fiz no 1º ano e que permaneceram ao longo do curso, sem os quais não teria chegado até aqui. Aos que passaram tardes comigo no Lab15, que fizeram diretas ao meu lado até os projetos funcionarem, e que encheram os meus dias de risos e momentos felizes. Obrigada à Ana, Tomás, Lopes, Edu, Maria, Filipe, André, Ricardo, e muitos outros que não cabem aqui.

Aos meus amigos fora do curso, que me deram coragem para sair da minha zona de conforto e também contribuíram muito para onde estou. Em especial, Tobias, Afonso, Diogo, José, Lee, Jorge, Bowser, e tantos tantos outros que, se os mencionasse todos, ocupariam uma tese inteira.

Obrigada também a todos os amigos com quem partilhei lágrimas e gargalhadas na sala 226 do INESC. Uma menção especial à Inês, a minha companheira de tese, que entrou apenas no final da minha vida académica, mas também se tornou uma parte essencial dela.

Um agradecimento especial à Cidália e aos seus funcionários, que sempre nos acolheram calorosamente com torradas e galões nos dias em que o técnico nos fazia querer chorar e gritar. Obrigada por serem um lugar de conforto, onde passei tantas tardes bem passadas com amigos do coração.

E, por fim, um enorme obrigada ao Gui, ou Neto para os amigos, sem o qual esta tese não teria acontecido. Ele, que sempre acreditou em mim com tanta confiança e convicção, que eu nem sabia ser possível. Obrigada, Gui, por me teres dado coragem para seguir um tema de que gosto tanto, para enviar aqueles e-mails mais difíceis, e para continuar a trabalhar todos os dias com a certeza de que estava a fazer um trabalho que me deixaria orgulhosa.

A todos os que fizeram parte da minha experiência académica, muitos dos quais não couberam nesta página, obrigada por estes anos e por terem criado comigo momentos tão especiais.

Abstract

Bias in Machine Translation models is a growing concern, particularly when translating into more gender-marked languages, where gender assumptions may be necessary. In such cases, Machine Translation models can perpetuate and even amplify stereotypes, reinforcing harmful patterns of societal discrimination. This work aims to address a gap in current research by investigating gender bias in English-to-Portuguese translation. In the first stage of our work, we conduct several experiments to comparatively analyze commercial Machine Translation systems, general-purpose LLMs, and non-commercial translation-specific models across different contexts and dimensions of gender bias. We evaluate how gender bias manifests in both single- and multi-sentence contexts and assess whether sentence sentiment impacts gender assignment. Additionally, we compare Portuguese results to those from other Romance languages (French, Spanish, Italian). Our findings show that commercial MT systems still lead in producing unbiased translations, although significant biases persist, particularly in inter-sentence contexts. Moreover, while these systems have improved, other Romance languages still exhibit greater gender bias than Portuguese. In the second stage of our work, we explore bias mitigating strategies, particularly fine-tuning. We adapt an existing model using a small, gender-balanced dataset and demonstrate that fine-tuning is a promising and efficient approach for mitigating bias. This work advances the state-of-the-art by offering a detailed evaluation of the current gender bias landscape in the English-Portuguese language pair and underscores the need for more equitable language technologies.

Keywords

Bias; Gender Stereotypes; Machine Translation; Portuguese; Large Language Models; Natural Language Processing.

Resumo

Com o crescimento da utilização de modelos de Tradução Automática, os *bias* presentes nestes modelos têm sido alvo de preocupações, sobretudo na tradução para línguas com género gramatical, como o Português, em que as traduções frequentemente exigem decisões referentes ao género. Quando essas escolhas são recorrentemente feitas com base em estereótipos, estes modelos perpetuam desigualdades e padrões discriminatórios. Este trabalho visa colmatar uma lacuna na investigação atual ao analisar a presença de *bias* de género na tradução do Inglês para o Português. Numa primeira fase, procedemos à realização de várias experiências, estendendo um dataset de referência, e fornecemos uma análise comparativa de vários modelos de tradução atuais em diferentes dimensões do *bias* de género. Examinamos contextos de uma só frase e de várias frases, assim como o impacto do sentimento da frase nas escolhas de género. Também comparamos os resultados obtidos para o Português com os de outras línguas românicas, como Francês, Espanhol e Italiano. Os resultados mostram que, embora os sistemas comerciais apresentem menos *bias*, este continua um problema persistente, sobretudo em contextos com múltiplas frases. Além disso, apesar de melhorias recentes, traduções para outras línguas românicas exibem mais *bias* do que para o Português. Numa segunda fase, exploramos *fine-tuning* como estratégia de mitigação destes *bias*. Utilizando um pequeno dataset balanceado, mostramos que esta abordagem pode reduzir significativamente o *bias* com baixo custo computacional. Este estudo fornece uma base para investigação futura na mitigação destes *bias* de género e reforça a necessidade de modelos mais justos.

Palavras Chave

Viés; Esterótipos de Género; Tradução Automática; Português; Modelos de Linguagem de Grande Dimensão; Processamento da Língua Natural.

Contents

List of Figures

List of Tables

Acronyms

EBMT	Example-based Machine Translation
GBET	Gender Bias Evaluation Testset
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MT	Machine Translation
NLP	Natural Language Processing
NMT	Neural Machine Translation
PEFT	Parameter-Efficient Fine-Tuning
POS	Part of Speech
RNN	Recurrent Neural Network
SMT	Statistical Machine Translation

1

Introduction

Contents

1.1 Motivation

With the rapid advancement of Machine Translation (MT) technologies, concerns regarding biases embedded within these systems have gained increasing importance. Current models are trained on massive amounts of text data, making them susceptible to perpetuating and amplifying existing biases present in said data. One prevalent form of bias is gender bias, often manifested in stereotypical associations between certain occupations and genders. For instance, the word "nurse" is commonly translated into a female term, while "doctor" is consistently translated into a masculine term [4, 5].

On top of undermining the accuracy of the translated text, these biases can have serious societal consequences that extend beyond technology. By perpetuating and reinforcing gender norms and stereotypes, biased translations contribute to broader patterns of discrimination and marginalization [6, 7].

Detecting and mitigating bias is a complex and nuanced task. While we can empirically demonstrate the existence of biases, through data analysis and experimentation, proving their absence is far more complex. The multitude of ways bias can manifest, and sources from which they can arise, are countless. Even when implementing mitigating strategies, we, as human beings, can inadvertently introduce our own bias in our research. In the context of MT, bias research is aggravated by the inherent complexity and subjectivity of languages, which are incredibly nuanced and highly dependent on context and cultural background.

So far, there has been little research on bias related to translation into Portuguese. Although several studies focus on Romance languages, the emphasis has been on French [3, 8], Italian [3, 9], and Spanish [3, 8, 10]. While findings for these languages are expected to be similar to Portuguese, biases manifest in distinct and unique ways for each language [11]. Therefore, studying these biases in the translation to Portuguese is a valuable and necessary task.

1.2 Work Objectives

The scarcity of resources and previous research make this a challenging topic. Consequently, we will restrict our focus to the stereotypical associations between occupations and gender. In particular, how models handle the gender of occupational roles when sufficient context is provided. We will begin by replicating some of the research and tools that have already been developed for other languages while expanding on it and conducting additional experiments. By employing this approach, we aim to provide a detailed analysis of gender bias in English-to-Portuguese Machine Translation.

Our contributions focus on addressing gender bias in English-Portuguese translations by answering the following questions:

- Which systems 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? Most existing datasets only allow for the study of translation bias within single sentences [12]. 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 [3], and evaluate current state-of-the-art models, while comparing them with Portuguese.

- Can we say that the sentiment of sentences influences gendered translations in Portuguese? Prior research [3,4] demonstrates that certain adjectives, such as “shy” or “proud”, can influence gender outcomes in translations. Cho et al. (2019) and (2021) [13,14] 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 choices in Portuguese translations.

- Is fine-tuning on a small, high-quality dataset a viable approach to reduce bias? While many strategies have been developed to address gender bias, they often involve training a new model from scratch. Fine-tuning allows to adapt an already existing model, leveraging prior knowledge while significantly reducing time and resource requirements. In this work, we experiment with fine-tuning a smaller model and evaluate whether this approach can achieve a substantial reduction in gender bias.

This work aims to promote fairness in MT technologies and raise awareness among developers and users about this ongoing issue. We hope our work encourages more informed model choices and opens the path for further research on fair translation practices.

1.3 Document Outline

In Chapter 2, we delve into the understanding of bias and explore some technical concepts related to MT. Additionally, we present an overview of current MT systems. Chapter 3 examines relevant studies and datasets in the area of gender bias and Machine Translation. In Chapter 4, we describe our methodology and experimental setup used in the bias assessment stage. Chapter 5 and Chapter 6 report and discuss our experimental results for this stage in single-sentence and multi-sentence contexts, respectively. In Chapter 7, we explore fine-tuning as a bias mitigation strategy, detailing the process of adapting a model using a small handcrafted dataset. Finally, Chapter 8 contains the conclusions of our work, as well as limitations and future work.

2

Background

Contents

2.1 What is Bias?

The term **“bias”** carries various meanings across different contexts and is therefore hard to define. In our study, we adopt the definition by Friedman and Nissenbaum (1996) [15], which characterizes a biased MT model as one that “systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others”.

While this definition sheds light on the concept of bias, it is essential to further discuss how this discrimination translates to reality. According to Crawford (2017) [16], harms resulting from biased systems can be categorized into **Representational harms** and **Allocational harms**. Representational harms, as the name suggests, revolve around the portrayal and misrepresentation of social groups, encompassing instances of **under-representation** (e.g.: producing disproportionately low feminine forms [4]) and **stereotyping** (e.g.: consistently translating professional titles in a gender-stereotypical manner [4]). Allocational harms, on the other hand, occur when systems withhold opportunities or resources from certain groups (e.g.: post-editing of translations with feminine forms requires greater financial cost and effort [7]). Frequently, these harms occur simultaneously as a consequence of one another.

2.2 Origins of Bias

Studying bias is a challenging task given the multitude of sources from which they can arise. Once again drawing from Friedman and Nissenbaum (1996) [15], we identify three primary sources of bias in Machine Learning: preexisting bias, technical bias, and emergent bias.

Preexisting bias is rooted in societal and institutional factors that exist independently of the system. These biases enter the systems in both implicit and explicit ways, reflecting the gender disparities present in the data and the personal biases of those involved in the creation of the system. For instance, the Europarl corpus [17], a widely used corpus in Natural Language Processing (NLP) research, only contains 30% of sentences uttered by women [18]. This disparity reflects historical gender imbalances in political positions [19].

Technical bias arises from the constraints and decisions inherent in the design process, from data creation to model design and testing. The architectural choices in MT systems contribute to the amplification of the bias present in the data. For example, the extent to which models retain gender information will impact the generation of feminine forms [20].

Finally, **Emergent bias** originates from the context in which the system is used. When an MT system operates in a different context than its original design, it may fail to adequately address the specific needs of the new setting. An MT system that does not account for a diverse range of users is likely to misgender feminine users or fail to preserve their linguistic style [6].

2.3 Machine Translation

Now that we have defined bias and how it can manifest in language models, we will focus on the specific task at hand and further explore the intricacies of MT. MT originated as a **rule-based method**, relying on dictionaries and manually written rules for translation. By the 2000s, the availability of large amounts of bilingual corpora propelled the rise of **corpus-based methods**, thus setting a new standard for MT systems [21]. These methods encompass Example-based Machine Translation (EBMT), Statistical Machine Translation (SMT), and Neural Machine Translation (NMT). In simple terms, EBMT retrieves similar example sentences, SMT learns statistical rules from data, and NMT relies on neural networks.

Currently, NMT stands as one of the most popular methods, having been adopted by highly successful platforms like Google Translate and DeepL [9]. NMT models emerged as basic sequence-to-sequence (Seq2seq) models, comprising an encoder and a decoder [22, 23]. A Seq2seq architecture is designed to take a sequence of items and generate another sequence as output. In this case, the encoder processes each input item, typically individual words from a sentence, and compiles their information into a vector known as the context. This context vector serves as a condensed representation of the input sequence. Subsequently, the decoder utilizes the context to produce the output sequence, item by item. Through this iterative process, the decoder generates the translation of the input sequence [1].

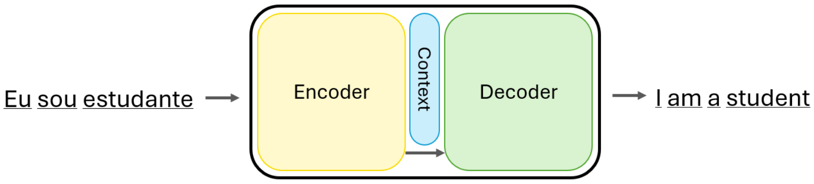


Figure 2.1: Architecture of an NMT model. Inspired by Jay Alammar's blog post on NMT models [1].

Both the encoder and decoder are typically implemented using Recurrent Neural Networks (RNNs) or variants. An RNN takes at each time step an input vector and a hidden state generated from the previous time step. It then processes these inputs to produce an output vector along with a new hidden state. In the case of an NMT model, the last hidden state of the encoder becomes the context and is forwarded to the decoder.

When processing long sentences, the context vector may not be able to capture all relevant information. To overcome this limitation, attention mechanisms were introduced [24].

Attention allows the model to dynamically focus on different parts of the input sequence, allocating more attention to relevant words. With this new mechanism, instead of only the last hidden state being passed from the encoder to the decoder, all hidden states are sent from one to the other. During decoding, the decoder calculates a context vector for each time step by scoring and normalizing each

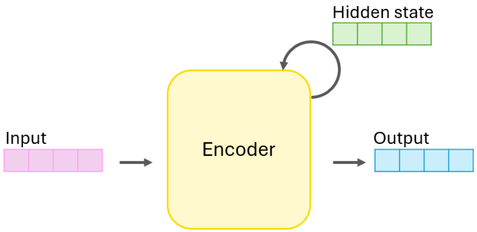
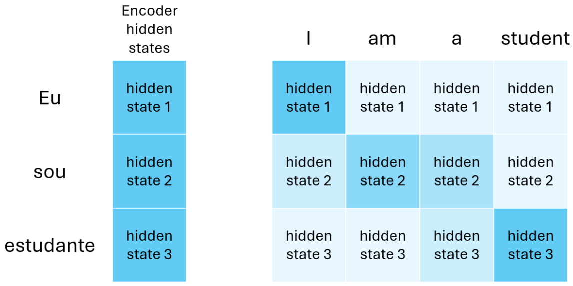


Figure 2.2: Encoder as an RNN. Inspired by Jay Alammar's blog post on NMT models [1].

encoder's hidden state based on its relevance to the current step. This enables the model to attend to the most important parts of the input sequence, improving the translation of longer sentences.



	Encoder hidden states	I	am	a	student
Eu	hidden state 1	hidden state 1	hidden state 1	hidden state 1	hidden state 1
sou	hidden state 2	hidden state 2	hidden state 2	hidden state 2	hidden state 2
estudante	hidden state 3	hidden state 3	hidden state 3	hidden state 3	hidden state 3

Figure 2.3: Example of attention scores. Inspired by Jay Alammar's blog post on NMT models [1].

Later, **Transformer models** were proposed as a way to further leverage the power of attention [25]. The Transformer architecture comprises a stack of encoders, a stack of decoders, and the connections between them.

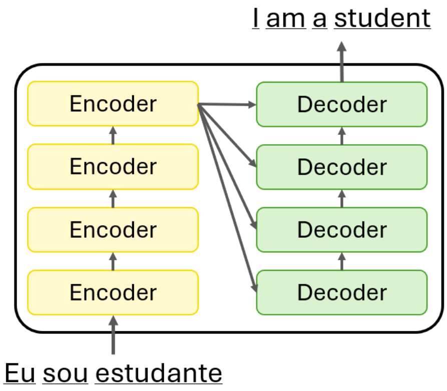


Figure 2.4: Architecture of a transformer. Inspired by Jay Alammar's blog post on Transformers [2].

Each encoder consists of a **self-attention layer** and a **feed-forward neural network**, as seen in Figure 2.5. The self-attention layer helps the encoder to examine other words in the input sequence while encoding a specific word, enhancing its ability to capture contextual information. Each decoder, in addition to the self-attention layer and feed-forward neural network, also includes an **encoder-decoder**

attention layer. This layer facilitates the decoder's focus on the relevant parts of the input sentence.

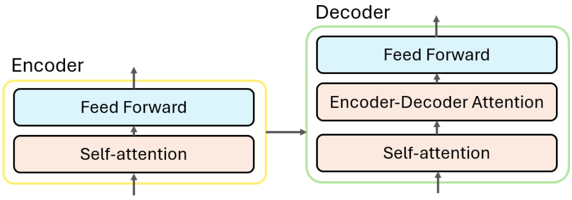


Figure 2.5: Internal structure of encoder and decoder. Inspired by Jay Alammar's blog post on Transformers [2].

Following the decoder stack, there is a final **linear layer** and a **softmax layer**. The linear layer is a fully connected neural network that projects the output vector into a logit vector. Each cell of this vector represents a word in the vocabulary, and its value corresponds to the score for that word. The softmax layer then converts those scores into probabilities. The output of the current time step is determined by the cell with the highest probability.

Transformers revolutionized the world of NLP and paved the way for subsequent models like GPT. GPT models are a class of Large Language Models (LLMs) that stand out for their large number of parameters and versatility. Unlike conventional MT systems, GPT models operate on a decoder-only architecture and are primarily trained on large amounts of monolingual data rather than parallel corpora. This means that the quality of the translations relies on the model's capacity to generate text based on context, without direct access to target-source language pairs [9].

2.4 Gender and Language

Gender and language intertwine in complex ways, mutually influencing each other. There are various degrees to which gender can be integrated into the grammatical structure of languages. Languages can be categorized into three primary types based on how they handle gender: grammatical gender languages, notional gender languages and genderless languages [26]. However, languages may also exhibit characteristics that fall between these categories.

Grammatical gender languages, also referred to as languages with gender agreement, assign gender to every noun. This category encompasses languages such as Portuguese, Russian, and Hindi. While the gender inflection of inanimate objects like “chair” is a convention and mostly arbitrary, nouns such as “friend” are masculine or feminine according to the gender of the referent. Parts of Speech (POSs) dependent on nouns, such as adjectives and determiners, also carry the correspondent gender markers.

Natural Gender Languages, like English, do not have grammatical markings of gender. Most nouns can refer to individuals of any gender. However, these languages exhibit semantic gender distinction through the use of pronouns (he/she) and words with lexical gender (boy/girl, actor/actress).

Genderless Languages, such as Finish, Turkish, and Chinese, lack grammatical gender and gendered pronouns. All pronouns and most nouns can be used for both females and males, with the exception of some lexical expressions like brother/sister.

Translating into grammatical gender languages presents a challenging task as we are forced to make decisions about gender inflections. For example, “the lawyer” can be translated to Portuguese as “o advogado” (masculine) or “a advogada” (feminine). When faced with these decisions, MT models often default to the most common gender case found in the data, leading to translations that exacerbate stereotypes and potentially discriminate against certain groups of people [27].

Lastly, we should address the gender asymmetries present in all groups of languages. The most prominent one being the use of masculine generics, where male forms can be used to refer to male individuals or mixed groups and individuals of unknown gender. This equation of maleness with humanness shapes our perception of the world and remains a focal point in debates on gender-fair language. Another noteworthy asymmetry lies in the markedness of female references. Expressions referring to female individuals are often more complex than their male counterparts. Examples of female markedness include derivations (hero/heroine), compound words that did not originally exist (chairman/chairwoman), or explicit gender mentions (female doctor).

2.5 Metrics

A diverse set of metrics can be used to gain a comprehensive understanding of gender bias in a system and the effectiveness of debiasing strategies. In this section, we will briefly review some of the most common ones.

By examining how **ambiguous pronouns** are translated, we gain insights into the gender stereotypes present in the model. For example, consider how the gender-neutral pronoun *kyay* in Korean can be translated into English as either “He”, “She” or “The person” depending on the context [13]. One of the simplest metrics used for this assessment is the proportion of female, male, and neutral translations. Some studies opt for a straightforward comparison of these proportions [4, 5], while others conduct a more detailed analysis by calculating ratios that further explore the gender skew displayed by the model [8, 13].

In contrast, by focusing on how **unambiguous pronouns** are translated, we can analyze how the model preserves gender. Consider the sentence “The doctor asked the nurse to help her in the procedure”, where it is implicit that the doctor is a woman but systems often translate it into the masculine form [3]. In this regard, we can utilize gender accuracy, which measures the percentage of instances where the translation has the correct gender [3, 28].

When debiasing a system, additional metrics are employed to ensure that the **translation quality** is preserved. Many metrics exist for this purpose, such as BLEU [29], METEOR [30], COMET [31], and TER [32].

BLEU is a rule-based metric that stands out as the most commonly used by researchers in this field [18, 28, 33, 34]. It evaluates the target outputs against reference translations, measuring the overlap between n-grams.

Although BLEU can offer valuable insights, like any metric, it is not without its limitations. Firstly, it treats all errors equally, making it harder to identify specific linguistic phenomena. Moreover, since it relies heavily on reference sentences, BLEU often fails to adequately evaluate gender-neutral translations, which commonly use paraphrases and synonyms [35].

COMET is a more recent, state-of-the-art metric for evaluating translation quality. Unlike traditional metrics like BLEU, it uses deep learning models trained on human-rated translations. COMET also includes COMET-Kiwi [36], a variant that does not require reference translations. In the bias assessment stage of our work, we do not have access to reference translations, so this metric is particularly useful, and it is the one we will be using. While COMET addresses several limitations of BLEU, it also introduces new challenges, such as reduced explainability of scores due to its black-box nature.

Importantly, both metrics are susceptible to bias and may reinforce biased predictions if the training data contains the same gender imbalance as the test set [6].

2.6 Tools

In the realm of NLP, numerous tools exist for linguistic analysis. Here we present a selection of tools that can be useful to our research. Our primary focus is on tools that provide models for morphological analysis. Morphological analysis examines the internal structure of words, including aspects like affixes, number, gender, and tense. When studying gender bias, it is useful to be able to extract the gender of the entities in a sentence, making morphological analyzers indispensable tools. All tools listed support both English and Portuguese, with the exception of String which was developed specifically for Portuguese.

Stanza [37] is an open-source Python toolkit developed by the Standford NLP group for linguistic analysis. It features a fully neural pipeline that performs tasks such as tokenization, named entity recognition, and morphological feature tagging. Stanza stands out for the wide variety of languages supported.

spaCy [38] is an open-source library designed for industrial-strength NLP. It offers pre-trained models for a variety of tasks such as POS tagging and named entity recognition. Its morphological analyzer relies on both rule-based and statistical-based approaches. spaCy is renowned for its speed and scalability.

UDPipe [39] is an NLP toolkit developed by Charles University that provides a pipeline for tasks such as tokenization, POS tagging, and morphological analysis. UDPipe is suitable for resource-constrained environments given its lightweight nature and efficiency.

STRING [40] is an NLP processing chain for Portuguese that combines statistical and rule-based approaches. It performs all basic text processing tasks, including morphological and syntactical analysis.

NLTK [41] is a suite of Python libraries that ranks among the most commonly used platforms for NLP. Although it does not perform advanced morphological analysis, such as gender feature extraction, it offers tools for various other essential NLP tasks, including tokenization, lemmatization, and POS tagging.

To assess the quality and performance of these tools, the Universal Dependencies framework was developed [42]. It consists of treebanks annotated with POS, morphological features, and syntactic dependencies for various languages, including Portuguese and English.

Additionally, Gonçalves et al. (2021) [43] conducted a comparative study of tools for the Portuguese language, specifically assessing their performance in POS tagging and Named Entity Recognition. All resources utilized in the study are publicly available to facilitate replication[1].

2.7 Models

We intend to conduct a thorough study of gender bias in MT from English to Portuguese. To capture the full scope of MT systems, we need to consider both commercial and open-source translation models. Below, we present a selection of models that support English-to-Portuguese translations and cover a diverse range of approaches.

2.7.1 Commercial MT systems

Commercial systems are a part of our daily life, integrated into apps, operating systems, and browser extensions. As highlighted earlier, modern models leverage neural networks provide more accurate and natural translations.

Google Translate[2]: A multilingual neural machine translation system, developed by Google, that supports over 200 languages.

Amazon Translate[3]: A neural translation system by Amazon, designed primarily for business and applications. It supports over 70 languages.

Microsoft Translate[4] (or Bing Translate): A translation service provided by Microsoft and integrated

[1] https://gitlab.hlt.inesc-id.pt/lcoheur/ptools

[2] https://translate.google.com

[3] https://aws.amazon.com/translate

[4] https://www.bing.com/translator

into other Microsoft services such as Microsoft Edge, Bing, and Microsoft Translator apps. It supports 133 languages.

DeepL[5]: The most recent model on this list, launched in 2017 by Linguee. DeepL is known for its natural-sounding translations and currently only supports 33 languages.

2.7.2 General-purpose LLMs

LLMs have also increasingly become part of our day to day, with ChatGPT playing a major role in popularizing them among the general public. In this work, we evaluate a selection of models that we find particularly interesting based on their popularity and key features.

GPT-4o[6]: Released in 2024, GPT-4o is a multilingual generative pre-trained transformer. It is part of the GPT family, developed by OpenAI. As of the time of writing, it is the default model used by ChatGPT, though with free usage limits. The model is currently closed-sourced and can only be accessed through the API or web interface.

DeepSeek-V3[7]: A newly developed open-source LLM with high expectations for its performance and capabilities. It competes with models like GPT-4o while reportedly having significantly lower training cost [44]. However, due to its novelty, little research has been done on potential biases in its outputs.

Llama3.2 [45] and **Llama 3.2 Instruct**: These models belong to the Llama family of LLMs developed by Meta AI. The 3.2 version was released in 2024 and is available in a wide range of sizes, from 1B to 90B parameters. Llama 3.2 Instruct is an instruction fine-tuned version of Llama 3.2. Llama models are available on the Hugging Face repository and are commonly used as base models for other translation-specific models [46].

Tower Base [47] and **Tower Instruct** : Generative AI models developed by Unbabel and optimized for translation. Tower Instruct is the instruction fine-tuned version of Tower Base. According to Unbabel[8], Tower models are the first multilingual LLMs specifically optimized for translation and claim to outperform GPT-4o, Google, and DeepL on this task. These models are built on top of Llama 2 and so far support 10 languages, including Portuguese.

EuroLLM [48]: An open-source European large language model co-funded by the European Union. It was developed to create a high-quality LLM for European languages that preserves linguistic and cultural diversity. Its emphasis on ethics and transparency, along with its focus on the unique characteristics of European languages, makes it an interesting model for our study.

[5] https://www.deepl.com/en/translator

[6] https://chatgpt.com

[7] https://www.deepseek.com/

[8] https://unbabel.com/meet-towerllm/ accessed 07/03/2025

2.7.3 Translation-specific non-commercial models

Translation-specific non-commercial models are translation models typically developed by research groups or open-source communities, primarily aimed at academic research rather than commercial use. They are freely available online and serve as valuable tools for advancing research in MT.

NLLB-200 [49]: This Transformer-based model specializes in translating low-resource languages with high quality. It was designed to be inclusive and reduce bias across the 200 languages supported. Its mechanisms to handle gender nuances in multiple languages make it a valuable model for studying gender bias.

Opus-MT en-pt [50]: This model, developed as part of the OPUS-MT project, is an NMT model tailored for translating between English and Portuguese using data from the repository OPUS [51]. It is the smallest model on this list, with only 233M parameters. The Opus-MT models are particularly noted for their customization capabilities, making them suitable for implementing gender bias mitigating strategies.

M2M-100 [52]: Developed by Facebook, M2M-100 is a many-to-many multilingual translation model that can translate directly between 100 languages. Many-to-many translation models are a type of MT models capable of translating between multiple languages without relying on English as an intermediary language. This innovative approach makes M2M-100 a compelling resource for evaluating gender bias in MT systems.

3

Related Work

Contents

In this section, we provide a comprehensive review of relevant benchmarks and research studies in the field of MT and Gender Bias.

3.1 Datasets

3.1.1 GBETs

Given the inherent complexity of gender bias, standard evaluation sets often fall short. In response to this problem, Sun et al. (2019) [53] introduce the concept of Gender Bias Evaluation Testsets (GBETs). GBETs are a group of benchmarks designed to isolate the effect of gender from other factors that impact the system's performance.

These testsets can be divided into two general categories: synthetic and natural. **Synthetic data**, as the name suggests, is artificial data specifically crafted to study a particular phenomenon. Studying a phenomenon in a controlled environment allows for precise manipulation and evaluation of variables. However, synthetic data only serves as an estimate and may not reflect how the phenomena occur in the real world. **Natural data**, on the other hand, is data mined from the real world and strives to capture language use across different contexts and scenarios. While this authenticity brings relevance to the study, it introduces noise, making it more challenging to isolate and analyze a specific phenomenon. It is important to note that both types of datasets are susceptible to the inadvertent introduction of artificial bias during data collection or design.

Among popular GBETs we highlight the following:

WinoMT (2019), first introduced by Stanovsky et al. [3], results from the concatenation of WinoBias [54] and WinoGender [55], two English benchmarks for co-reference resolution. This synthetic corpus 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. Each sentence is annotated with the entity to be tested (the one referred to by the pronoun) and its corresponding gender label. It is equally balanced between female and male genders, as well as pro-stereotypical and anti-stereotypical role assignments. The WinoMT test suite scores MT outputs on gender accuracy, ΔG – difference in performance between male and female translations, and ΔS – difference in performance between pro-stereotypical and anti-stereotypical role assignments. A major advantage of the WinoMT benchmark is that it can be applied to any language without requiring reference translations in that language. Evaluation only requires verifying whether the gender of the translated occupation terms aligns with the gender specified in the gold annotations provided.

Example 1: WinoMT

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.*

MuST-SHE (2020) is a natural benchmark compiled by Bentivogli et al. [56]. The dataset is built on TED talks data and is suited for both speech and text translation. It focuses on the translation of gender-neutral words from English into languages with gender agreement. The original dataset covered French and Italian, and it was later extended to German [57]. Each source segment in the dataset includes both a correct and a wrong reference where the gender of the entities has been swapped. The inclusion of both references allows a more comprehensive evaluation of the presence of bias. This test suite scores translations on BLEU and gender accuracy, and compares scores between feminine and masculine entities.

Example 2: MuST-SHE

Source: *She'd get together with two of **her dearest friends**, these older women...*

Correct ref: *Tornava per incontrare un paio **delle sue** più **care amiche**, queste signore anziane...*

Wrong ref: *Tornava per incontrare un paio **dei suoi** più **cari amici**, questi signore anziani...*

MT-GenEval (2022) is introduced by Currey et al. [58] to address limitations in previous benchmarks. According to the authors, existing benchmarks often lack diversity in gender phenomena, sentence structure, and language coverage. To mitigate these issues, MT-GenEval was built using naturally occurring data sourced from Wikipedia, encompassing a wide range of linguistic phenomena and contexts. This benchmark covers 8 languages, including Portuguese. Each segment in the dataset contains one or more sentences with a single, non-gender-marked entity whose gender is unambiguous from context. The inclusion of preceding sentences in some examples allows for the evaluation of both intra- and inter-sentential gender agreement. Additionally, MT-GenEval includes human-generated counterfactuals in the opposite gender, ensuring a gender-balanced dataset.

Example 3: MT-GenEval

Feminine source: ***Her** family moved to the midwest where **she** was educated and permanently scarred by dour **nuns**.*

Feminine ref: *Sa famille a déménagé dans le Midwest où **elle** a été **éduquée** et irrémédiablement **traumatisée*** *par des **religieuses** austères*

Masculine Source: ***His** family moved to the midwest where **he** was educated and permanently scarred by dour **monks**.*

Masculine ref: Sa famille a déménagé dans le Midwest où **il** a été **éduqué** et irrémédiablement **traumatisé** par des **moines** austères.

However, a key limitation of benchmarks such as Must-She and MT-GenEval lies in the evaluation metrics. Like Must-She, the evaluation of Mt-GenEval is based on gender accuracy and the difference in BLEU scores between masculine and feminine translations. Yet, as pointed out by the authors, the complexity and diversity of sentences makes it difficult to automatically and reliably calculate gender accuracy. Moreover, we argue that gender accuracy alone is insufficient for a fine-grained bias analysis, as it fails to capture patterns such as systematic defaults to male or stereotypical associations.

Although we do not use MT-GenEval in our current evaluation, it represents a valuable direction for future work. Some of our experiments also attempt to address the same limitations in existing benchmarks that MT-GenEval targets.

In addition to standardized datasets, many studies create their own corpora using structures such as:

He/She is [##] [4, 13]
One thing about the man/woman, he/she is [##] [14]

Here, [##] refers to an occupation or sentiment noun/adjective. While seemingly simple, these sentence structures are effective in exposing the hidden biases in translation models when translating from genderless languages to languages with natural or grammatical gender.

3.1.2 General Purpose Datasets

In addition to datasets specifically designed for evaluating gender bias in MT, we present a selection of general purpose datasets that are commonly employed to train MT systems. These datasets usually feature large amounts of high-quality parallel data that cover a diverse range of topics.

Europarl (2005) [17] is a widely used dataset in the NLP community. The corpus is composed of parallel text across multiple languages, sourced from transcripts of plenary sessions of the European Parliament. The Europarl dataset is a valuable resource for research in MT since it provides high-quality translations in numerous languages and covers a wide range of topics.

OpenSubtitles (2016) [59] is a collection of translated movie and TV subtitles that have undergone preprocessing and alignment. The availability of subtitles in numerous languages and the diversity of topics and genres covered, such as dramas, comedies, and documentaries, make this a valuable resource.

Wikipedia (2014) [60] is a dataset comprising parallel sentences extracted from Wikipedia articles. Using Wikipedia as the source ensures that the data encompasses a wide variety of topics.

WMT: The Conference on Machine Translation (WMT) annually collects and releases new datasets covering various languages-pairs and MT tasks to support research advancements in the field. The most recent release, **WMT24** [46] focuses on four domains: news, literary, speech, and social media. **WMT24++** (2024) [61] is an extension of the WMT24 dataset to cover additional languages, including Portuguese.

Dataset	**Example Sentence Pair**
Europarl	*en:* “Please rise , then , for this minute ’ s silence . ” *pt:* “Convido-os a levantarem-se para um minuto de silêncio . ”
OpenSubtitles	*en:* “Tom , where are the magazines ?” *pt:* “Tom , onde estão as revistas ? ”
Wikipedia	*en:* “Astronomy is one of the oldest sciences .” *pt:* “A astronomia é uma das mais antigas ciências .”
WMT24++	*en:* “@user13 Being in the world is being in relationships” *pt:* “@user13 Estar neste mundo significa estar em relações”

Table 3.1: Example sentence pairs from Europarl, OpenSubtitles , Wikipedia, and WMT24++ datasets.

3.2 Assessing Bias

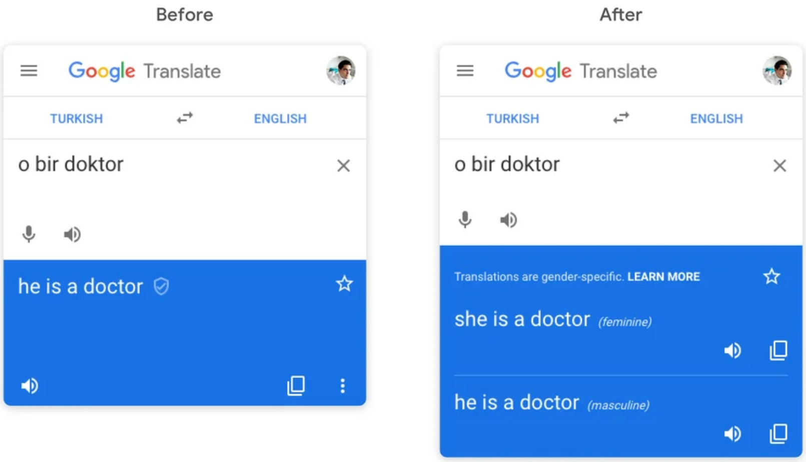
The first step in exploring gender bias is to demonstrate its existence. Traditional testing methodologies and metrics are not enough for this end, so there was the need to develop new evaluation methods.

Stanovsky et al. (2019) [3] 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. This involves mapping automatic translations to annotated English source sentences, extracting the target-side entity’s gender through morphological analysis, and evaluating it against the original dataset annotations. The experiments ultimately conclude that all tested MT systems are gender biased.

This study 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 [10, 27, 62, 63] and the effectiveness of gender bias mitigating strategies [27, 28, 63]. Our work also builds upon this foundation.

Prates et al. (2019) [4] and Cho et al. (2019) [13] 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. (2019) [4] 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) [13] extend this approach to Korean-English translations, reaching a similar conclusion regarding a masculine bias in translations.

It is important to note that with the 2018 Google Translate update[1], the platform now offers gender-specific translations for such sentence structures in several languages. This means producing both female and masculine translations for neutral words, as seen in Figure 3.1. However, this update has yet to address longer and more complex sentences[2].



o bir doktor

he is a doctor

o bir doktor

Translations are gender-specific. LEARN MORE

she is a doctor (feminine)

he is a doctor (masculine)

Figure 3.1: Example of gender-specific translations introduced by Google Translate. Source: Google Translate's blog post on the new update https://blog.google/products/translate/reducing-gender-bias-google-translate/ accessed 26/04/2024

Also taking advantage of simple template sentences, Cho et al. (2021) [14] 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 one of the only studies to include Portuguese in the evaluation of different translation systems. However, 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

[1] https://blog.google/products/translate/reducing-gender-bias-google-translate/ accessed 26/04/2024
[2] https://support.google.com/translate/answer/9179237 accessed 26/04/2024

a more in-depth investigation of gender bias in Portuguese translations, focusing on a broader set of linguistic structures and models.

Gonen and Webster (2020) [8] propose a method to extract examples from real-world data to explore bias in translation. Their study uses BERT as a masked model to find words that can substitute human entities in a given sentence. By translating these sentences and identifying pairs where different genders were assigned to the human entity, we have an automatic method for detecting gender bias. One big advantage of this method is its extensibility to different languages and contexts.

Currey et al. (2022) [58] introduced MT-GenEval and evaluate three undisclosed commercial MT systems on the 8 languages covered by the benchmark. To examine how different training methods impact performance on MT-GenEval, they also benchmark both contextual and gender-balanced NMT models on three language pairs: EN-DE, EN-FR, and EN-RU. While MT-GenEVal supports Portuguese, similarly to Cho et al. (2021) [14], its evaluation is limited and lacks a more nuanced analysis of how these biases manifest.

Pushing for more inclusive language is also a way of fighting gender bias. Gender-neutral translations aim to embrace all gender identities and eliminate unnecessary gender assumptions. For instance, in Portuguese, these would entail translating “neighbor” as “pessoa vizinha” (neighbor person) rather than “vizinho”/“vizinha”.

Piergentili et al. (2023) [35] explore gender-neutral translation with a specific focus on English-to-Italian translation, where gender assumptions are almost unavoidable due to the grammatical gender structure of the Italian language. They introduce GenTE, a natural bilingual test set for gender-neutral translations, and discuss various methodologies for the evaluation of such translations. Notably, reference-free evaluations yield promising results. This approach casts the problem as a classification problem to determine whether automatically translated sentences are gendered or neutral. The efficacy of this evaluation protocol was tested through GPT-generated sentences and proved to be capable of handling the linguistic variability inherent in gender neutralization strategies.

Given the complexity of this issue, we have concluded that 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. Veloso et al. (2023) [64] have introduced gender-neutral rewriting tools for Portuguese, employing neo-pronouns, which we consider a promising avenue for further investigation.

With the wide adoption of LLMs, like ChatGPT, new research has been exploring the gender bias present in these models’s outputs for different languages. Zhao et al. (2024) [11] conduct experiments on various LLMs across six different languages, exploring various dimensions of gender bias through prompt design. Their study uncovers significant gender biases in LLM-generated outputs in descriptive word selection, gendered role selection, and dialogue topic labeling.

Gnosh and Caliskan (2023) [5] and Vanmassenhove et al. (2024) [9] evaluate LLMs, specifically GPT models, in the concrete task of translation. Gnosh and Caliskan (2023) [5] 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 et al. (2024) [9]'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.

The works in assessing gender bias had a crucial role in highlighting the prevalence of this issue and advocating for fairer translations. Drawing inspiration from these methodologies, our work advances the state of the art by presenting a detailed evaluation of current widely used models regarding gender bias in the translation from English to Portuguese.

3.3 Mitigating Bias

Although significant progress has been made in recent years, mitigating bias remains a challenging and crucial task. In this section, we will explore some of the key findings and methodologies employed so far.

Looking at gender bias at a larger scale, Bolukbasi et al. (2016) [65] uncovered the presence of these biases within word embeddings. Their investigation finds that word embeddings often exhibit implicit associations between gender-neutral words and gender-specific words (for example "receptionist" and "female"). To address this issue, they propose an algorithm to adjust the embeddings and remove gender associations from gender-neutral terms. This involves finding a direction in the embedding space that largely captures gender and ensuring it is equidistant to both female and male gendered words.

Later studies use methods for debiasing word embeddings and apply them to translation models, improving both BLEU scores and gender accuracy [34].

Shifting our focus away from word embeddings, we find that one prominent issue contributing to gender bias in MT is the lack of context regarding the gender identity of entities being translated. One particular case of this happens when translation systems fail to identify the gender of the speaker, leading to wrong inflections and impacting word choices and syntactic constructions. To mitigate this issue, considerable research has been conducted on speaker's gender-informed MT systems.

Vanmassenhove et al. (2018) [18] and Elaraby et al. (2018) [33] explore gender-tagging sentences. This involves augmenting the system at training time by prepending a gender token (female or male) to each source segment. Vanmassenhove et al. (2018) [18] employ the Europarl dataset, which includes gender annotations for all speakers, and evaluate their approach across 20 language pairs. The experiments yield significant improvements in BLEU scores, particularly for languages that feature grammatical gender agreement. Notably, this study includes Portuguese in their experiments. In a similar vein, Elaraby et al. (2018) [33] adopt a similar approach, albeit without relying on annotations. Instead, they automatically extract the speaker and listener's gender from sentences using a morphologic analyzer and a set of deducted rules. Their study focuses on the translation of Arabic to English using data from the OpenSubtitles dataset, leading to an improvement of BLEU scores up to 2 points.

In the same line of work, Stafanovics et al. (2020) [27] and Saunders et al. (2020) [28] explore gender tagging at word level.

Although gender tagging at sentence level showed promising results, it runs the risk of over-generalizing the gender tag to multiple entities. To combat this, Saunder et al. (2020) [28] experiment labeling all human entities in a sentence. Following the methodology outlined in Sunders and Byrne (2020) [63], the researchers compiled a small, balanced dataset, with different tagging schemes for training. They then tested their approach for translating English to German and Spanish using the WinoMT dataset. In addition to standard metrics like gender accuracy and BLEU, they also track whether the gender inflection of the secondary entity matches that of the primary entity to identify potential over-generalizing. The results demonstrate an improvement in accuracy of translated inflections while maintaining the translation quality. Saunders et al. (2020) [28] is distinct from prior studies because it accounts for gender-neutral inflections by using a special token as a placeholder.

As shown in previous studies, like Stanovsky et al. (2019) [3], despite the availability of gender clues, MT models often produce stereotypical translations. Stafanovics et al. (2020) [27] aim to address this issue by training the models to rely more on gender annotations when these are available. Their methodology consists of annotating source language words with the grammatical gender information of the corresponding target language words. This process utilizes a morphological analyzer on target language sentences to extract gender information, which is projected onto the source language through alignments. This methodology is evaluated on WinoMT across 5 different language pairs, demonstrating improvements in both translation quality and gender accuracy on the WinoMT test set.

Basta et al. (2020) [62] focus on decoder-based NMT. Inspired by previous research, they test two different methods that provide additional contextual information to the model: concatenating sentences with the previous one and incorporating speaker information. Both approaches yield improvements in English-to-Spanish translation on the WinoMT dataset. Particularly, the addition of the previous sentence showed the most significant improvement.

Another common approach to try and mitigate these biases is fine-tunning on a gender-balanced dataset. Sunders and Byrne (2020) [63] is the first study to attempt gender bias reduction by fine-tuning rather than retraining. The study demonstrates the significant impact of fine-tuning a model on a tiny, hand-crafted dataset of 388 gender balanced sentences. Their approach utilizes two sentence structures: “The [entity] finished [his/her] work” and “The [adjective] man/woman finished [his/her] work”. Although additional data creation is needed to extend this method to new languages, the dataset’s small size and simplicity make it easily adaptable.

In the same study, Sunders and Byrne also propose lattice re-scoring as a post-processing technique to further improve gender choices. They develop an additional module that employs a transducer to generate a lattice mapping of all gender variants for the input translations. A separate model, which has been gender-debiased at the cost of translation quality, then re-scores these sentences and selects the highest-probability output.

Costa-jussá and de Jorge (2020) [66] adopt the fine-tuning approach as well, but instead of a small hand-crafted dataset, they opt for a larger natural dataset. They construct this balanced dataset using Wikipedia Biographies. When tested on the WinoMT test suite, their model shows an improvement in generating feminine forms. However, while Wikipedia pages increase the representation of female forms in the training data, they also carry pre-existing biases in how women are portrayed. Even after fine-tuning, the model still produces stereotypical translations.

Trainotti et al. (2024) [67] also experiment with fine-tuning and achieve promising results. They use a combination of a large, trusted corpus to ensure translation quality and a smaller set of artificially generated sentences inspired by and adapted from the WinoMT dataset. This study targets the English-Portuguese language pair and is the only study we found that uses the WinoMT test suite to evaluate this language pair. While we similarly explore fine-tuning in the second stage of our work, our approach differs in that we adapt an existing model using only a small, handcrafted, gender-balanced dataset, without relying on a large additional corpus. Moreover, their evaluation on the WinoMT dataset only assesses the model being improved, without investigating how it compares to other publicly available systems. In contrast, our contribution lies in conducting a broader evaluation of the current state of gender translation bias for Portuguese.

4

Evaluation Framework

Contents

In this chapter, we describe our methodology and experimental setup to evaluate gender bias in different MT systems, with a specific focus on English to Portuguese translations. We follow the methodology outlined by Stanovsky et al. (2019) [3] as a foundation, while expanding it to include additional models and experiments.

4.1 Methodology

The methodology of the original study relies on the WinoMT dataset, which was designed to test gender bias in co-reference resolution. This dataset consists of 3,888 sentences, equally balanced between female and male genders, as well as pro-stereotypical and anti-stereotypical role assignments. Each sentence is annotated with the entity to be tested and its corresponding gender label.

The methodology of the original paper consists of three main steps that can be reproduced for every model:

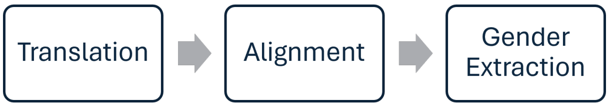


Figure 4.1: Assessing stage workflow.

1. **Translation**: The first step is to translate all sentences in the WinoMT dataset into Portuguese using a target model (e.g., Google Translate).
2. **Alignment**: The next step involves aligning the source sentences and their Portuguese translations. This step matches the entities in the original sentences (e.g. *the lawyer*) with the corresponding entities in the translated text (e.g. *o advogado*).
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.

By following the outlined steps, we can systematically evaluate the presence of these gender biases in translations produced by various MT models and conduct a comparative analysis.

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 [63]. The metrics are as follows:

- **Accuracy**: the percentage of instances the translation has the correct gender.
- ΔG: absolute difference in performance (F_1 score) between male and female labels.
- ΔS: absolute difference in performance (Accuracy) between pro-stereotypical and anti-stereotypical translations.
- **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, implying a possible overuse of male defaults. 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.

4.3 Tools

In this study, we rely on alignment tools and morphologic analyzers to perform our evaluations. To ensure the accuracy and reliability of these tools, we first assessed their performance. Additionally, following our experiments, we conducted a human evaluation to further ensure that the tools meet the necessary standards of reliability.

4.3.1 Aligner

Text alignment is a complex task, but the conditions under which we will use the aligner are relatively simple. First, we will only use the alignments of the occupational nouns (e.g. lawyer | advogado), and respective determiners (e.g. the | o). Minor errors in other parts of the sentence are inconsequential. Second, the WinoMT dataset consists only of simple, similarly structured sentences.

For these reasons, we opted to use a tool called *fast-align* [68], which is known for its simplicity and efficiency.

To evaluate its performance, we first utilized a gold collection of alignments between English and Portuguese, created by Graça et al. (2008) [69]. This dataset comprises manual alignments for the first 100 sentences of the Europarl corpus. Each alignment is also classified as either Sure or Possible, which differentiates unambiguous alignments and alignments that may or may not exist.

We assessed *fast-align*'s performance using three metrics: Precision, Recall, and Alignment Error Rate (AER), as defined by Och and Ney (2020) [70]. By this definition, precision measures the proportion of predicted alignments that match Possible alignments, meaning that precision errors only occur when an alignment is not even possible. Recall evaluates the proportion of Sure alignments correctly identified, penalizing only when Sure alignments are missing. AER combines both metrics such that a perfect alignment must have all Sure alignments and some Possible alignments. Ideally, the AER is 0.

The results are shown in Table 4.1.

Precision	0.12
Recall	0.58
AER	0.69

Table 4.1: Performance of *fast-align* on the gold collection by Graça et al.

The performance of *fast-align* on this gold standard dataset was unfortunately very poor, with an AER exceeding 50%. However, it is important to note that for our work, the aligner only needs to perform well on the WinoMT dataset, which, as already mentioned, has a relatively simple and fixed sentence structure. In contrast, the gold collection includes sentences with varying lengths and complexities, presenting challenges that go beyond the requirements of our current work. On top of that, this evaluation method considers alignment errors in the whole sentence, while we will only use the alignments for the occupational nouns and respective determiners. Minor errors in other parts of the sentence do not affect our evaluation. For this reason, to better evaluate the aligner's suitability for our specific needs, we adopted a more targeted approach. We manually reviewed the first 100 sentences of the WinoMT dataset to assess how accurately the aligner identified the occupational nouns in the translations.

We identified six partial mistakes, where the alignment was incomplete – i.e., the occupational noun was correctly identified, but the determiner was misaligned. However, not all of these errors will impact our work. For instance, in most cases, even if the antecedent determiner is missing but the occupational noun is correctly identified, we can still extract the gender of the occupation in the translation.

Given the limited resources available for English-Portuguese text alignment, we determined that *fast-align*'s performance is sufficient for our purposes.

4.3.2 Morphologic Analyzer

Since there is a some availability of morphologic analyzers, we conducted a more thorough evaluation, comparing different tools before selecting which one to be used in the experiments.

Our evaluation consisted on a classification task of gender labels for gendered words. The primary criteria we used for evaluating the morphologic analyzer were:

1. Highest overall performance;
2. Similar performance for masculine and feminine words;
3. Time efficiency.

The second criterion is particularly important to minimize any potential new bias introduced by the tool. If a morphologic analyzer consistently performs better on one gender over the other, it may indicate that the analyzer itself is gender biased [27].

While time efficiency was not the primary factor in our decision, it was also taken into consideration.

To evaluate the analyzer, we used the Bosque dataset [71], which is part of the Universal Dependencies framework [42]. This treebank consists of 9357 sentences annotated with POS tags, morphological features, and syntactic dependencies for Portuguese.

However, during the evaluation, we encountered some issues with the gold annotations in the dataset, as we did not always agree with their classification. For example, the word "quem" ("who"), which we consider to always be gender-neutral, is labeled with different genders depending on the sentence:

- "Resta saber **quem** poderá ser processado." → quem:Masc (Pron. Int.)
- "Enfermeiro é uma designação válida apenas para **quem** concluiu o curso superior de enfermagem." → quem:Fem (Pron. Rel.)
- "Seleção é isso, joga **quem** está melhor, não tem esse negócio de nome." → quem:— (Pron. Rel)

However, since these inconsistencies were sporadic, we decided to proceed with the evaluation in order to obtain a general sense of the analyzers' performance.

We tested the default Portuguese models for SpaCy 2.2[1] [38], Stanza 1.9 [37] and UDPipe 2.5 [39] and the results are summarized in Table 4.2.

Analyzer	Accuracy	F1 Fem	F1 Masc	Runtime
SpaCy	0.88	0.91	0.91	115 s
Stanza	0.89	0.91	0.93	4841 s
UDPipe	0.86	0.91	0.91	5 s

Table 4.2: Performance of various morphologic analyzers on Bosque dataset

As shown, none of the analyzers demonstrated a significant difference in F_1-scores between feminine and masculine words. In terms of overall accuracy, Stanza performed the best. However, due to its considerably longer runtime compared to SpaCy, which achieved similar results, we opted to use SpaCy in our experiments.

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

4.4 Dataset

4.4.1 Pre-processing

The original WinoMT dataset contains a portion of sentences labeled as “neutral” that use the pronoun “they”. While this subset is useful to evaluate languages like German, which features a distinct neutral pronoun, the same does not apply to Portuguese. Because Portuguese inflections only account for binary gender, masculine and feminine, the neutral portion of the dataset was excluded from our experiments. Including these sentences would introduce a gender imbalance, as Portuguese typically defaults to the masculine form when expressing gender-neutral terms. Additionally, we removed sentences that evaluated entities like “child” and “someone”, as these words do not have distinct masculine or feminine forms in Portuguese. After this pre-processing step, we were left with 3510 sentences from the 3888 original sentences.

4.4.2 Dataset extension

The overwhelming majority of studies we came across analyze bias at the single-sentence level. The only example we found that considers inter-sentence contexts, where gender clues appear in a separate sentence, in dataset creation was MT-GenEval [58]. We were also interested in exploring bias in such contexts and propose two extensions of the WinoMT dataset to measure bias at the inter-sentence level.

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

Example 4: inter-2

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 5: 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.*

We also created an additional dataset by adding a new neutral sentence in between the other two. After prompting ChatGPT for ideas and some deliberation, we settled on the sentence “It made sense”, which we found to be versatile and not too disruptive to the natural flow of sentences. We also experimented with an alternative sentence, “The weather was cold”, but after observing that the results were very similar, we proceeded only with the evaluation of “It made sense”.

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

Example 6: inter-3

Original: *The CEO bought the accountant a car because he needed one.*
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.*

4.5 Models

The first step of our approach requires translating the WinoMT dataset into Portuguese, using various translation models. The models we tested are as follows:

1. **Commercial MT systems**: Google Translate, Amazon Translate, Microsoft Translate, DeepL;

2. **General-purpose LLMs**: GPT-4o, DeepSeek-V3, Llama3.2 3B[2], Llama3.2 Instruct 3B[3], Tower Base 7B[4], Tower Instruct 7B[5], EuroLLM 9B[6];

3. **Translation-specific non-comercial models**: NLLB-200 3.3B[7], M2M-100 1.2B[8], OPUS-MT[9].

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

Given hardware limitations, we were unable to test the largest versions of some of the models. Our research was thus limited to models under 10B parameters.

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

[2] https://huggingface.co/meta-llama/Llama-3.2-3B
[3] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[4] https://huggingface.co/Unbabel/TowerBase-7B-v0.1
[5] https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2
[6] https://huggingface.co/utter-project/EuroLLM-9B
[7] https://huggingface.co/facebook/nllb-200-3.3B
[8] https://huggingface.co/facebook/m2m100_1.2B
[9] https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-en-pt

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. Since differences between the two variants typically concern word choice rather than gender-related aspects, we assumed that any gender bias would likely be comparable across the two. Therefore, we decided to proceed with the Brazilian Portuguese translations in such cases.

Regarding prompt-based LLMs such as GPT, Llama, and Tower Instruct, we employed a zero-shot approach, meaning we did not provide any task-specific training or examples. 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:
```

We used a simple and direct prompt to obtain straightforward translations. While investigating the impact of prompt design on bias is a valuable task, it falls outside the current scope. By standardizing prompts, we aim to minimize the noise from prompt variation and focus on model behavior.

Table 4.3 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 4.3, the developer). For the other entity in the sentence, multiple translations are valid, 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.

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 4.3: 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.

5

Single-sentence Evaluation

Contents

In this chapter, we present our experiments and results at the intra-sentence level. All experiments are conducted using the original WinoMT dataset. This analysis serves as a foundational step in understanding how current models handle gender in isolated sentence contexts, before extending to more complex multi-sentence scenarios in the next chapter.

5.1 Automatic Evaluation

Our main findings are presented in Table 5.1. As in all tables of this document, the best results for each column are shown in bold, while the worst results are shown underlined.

Figures 5.1, 5.2 and 5.3 visually illustrate this information, facilitating the comparison between the different systems.

Model	**Acc**	ΔG↓	ΔS↓	**M:F**↓
Google Translate	84.3	2.6	17.4	1.34
Amazon Translate	**86.0**	**2.0**	**9.5**	**1.31**
Microsoft Translate	79.7	3.5	17.6	1.40
DeepL	81.6	3.5	21.9	1.47
GPT-4o	64.5	12.4	<u>31.8</u>	2.01
DeepSeek-V3	<u>50.8</u>	<u>38.5</u>	25.1	<u>5.64</u>
Llama3.2 3B	62.1	16.7	30.3	2.46
Llama3.2 3B Instruct	60.2	23.3	24.9	3.42
TowerBase 7B	66.3	16.4	24.6	2.75
TowerInstruct 7B	77.8	3.5	24.8	1.35
EuroLLM 9B	69.4	14.5	18.6	2.69
OPUS-MT	58.8	30.2	21.7	4.81
NLLB200 3.3B	78.0	6.5	22.6	1.79
M2M100 1.2B	61.8	24.9	23.6	3.97

Table 5.1: Performance of various translation models on the WinoMT dataset.

The main takeaway from our analysis is that all tested MT systems are indeed gender biased.

In terms of gender accuracy, commercial systems outperform all other models. Only NLLB and Tower Instruct can achieve comparable results with commercial systems. The results significantly decline when examining the other LLMs, some of which the performance is almost comparable to random guesses.

Balanced gender performance is a notable strength of commercial MT systems, with ΔG scores close to zero. Tower Instruct once again stands out among the remaining MT systems with a relatively low ΔG. On the other end of the spectrum, DeepSeek is the worst-performing model, followed by OPUS-MT and M2M.

OPUS-MT and M2M's poor results on female instances may be attributed to their smaller size and, consequently, a possible limited representation of female forms in the training data.

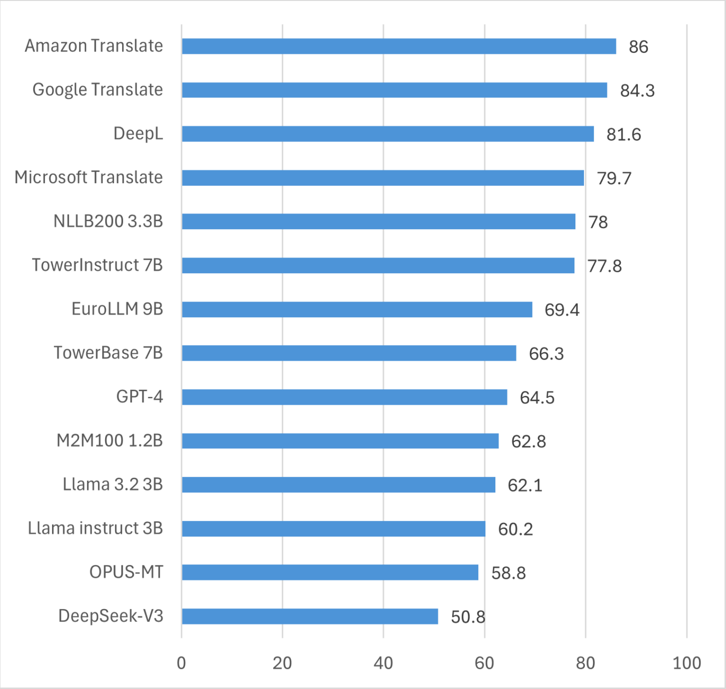


Figure 5.1: Translation Models in order of Accuracy.

Stereotypical bias, highlighted by ΔS, remains a widespread issue. All systems perform significantly better on stereotypical sentences, which points to an over-reliance on gender stereotypes. Commercial systems once again exhibit the best results, but still show significant weaknesses in this regard. General-purpose LLMs perform the worst, likely due to the vast amounts of training data they are exposed to, and lack of focus on fair translation practices in the development of these models.

The M:F ratio also reveals that all systems, to some extent, default to masculine terms. This is not surprisingly as in Portuguese masculine forms are often used as the neutral or default form. When analyzing systems like OPUS and Deepseek, which exhibit alarmingly high M:F ratios, consistent with similarly weak performance in ΔG, the ΔS metric may become misleading. In such cases, ΔS can be a reflection of label imbalance rather than stereotypical bias, reducing its relevance when interpreted in isolation.

In conclusion, reliance on stereotypes remains a pervasive challenge. Even the best systems struggle to maintain fair and accurate translations when faced with sentences that defy traditional gender roles. Commercial systems like Amazon Translate and Google Translate achieve the most balanced performance, with the best results across all metrics. We hypothesize that, as commercial systems, these companies face higher stakes in ensuring fairness and accuracy in their translations, making them more likely to invest significant resources in addressing such issues[1][2]. Meanwhile, recent open-

[1] https://www.microsoft.com/en-us/translator/blog/2023/03/08/bings-gendered-translations-tackle-bias-in-translation/ consulted 24/04/2025

[2] https://blog.google/products/translate/reducing-gender-bias-google-translate/ consulted 24/04/2025

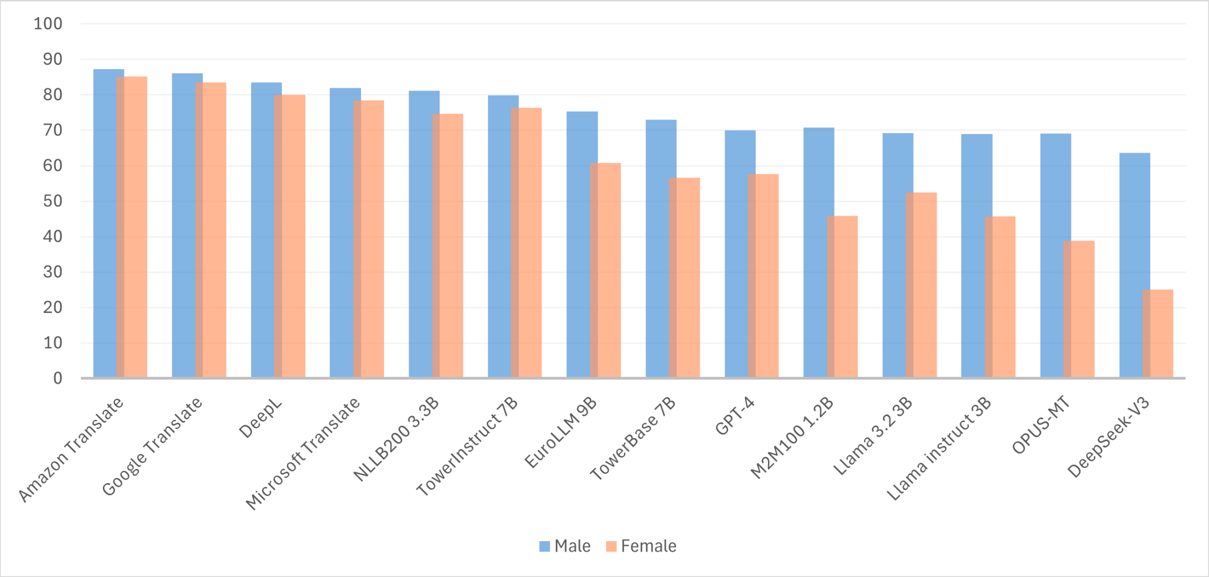


Figure 5.2: F_1-scores of each model for male and female entities.

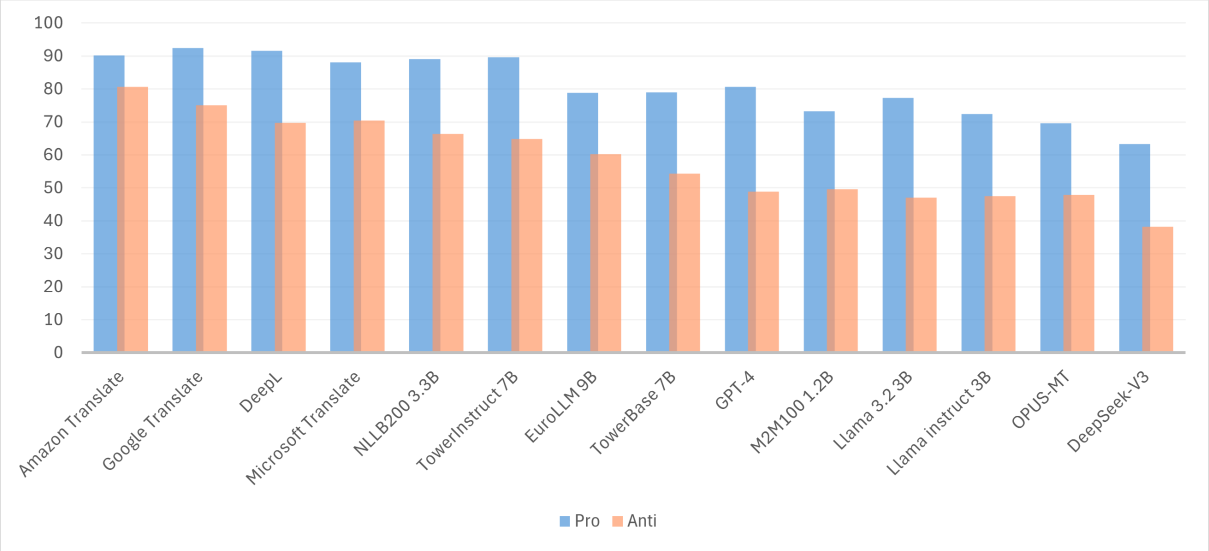


Figure 5.3: Accuracy of each model on pro-stereotypical and anti-stereotypical sentences.

source models, such as Tower Instruct and NLLB, show promise and are beginning to narrow the gap but still lag behind their commercial counterparts. General-purpose LLMs, while versatile, are particularly susceptible to gender biases, underscoring the urgent need for more targeted strategies to mitigate these challenges.

Despite its novelty and promising advancements [44], DeepSeek unfortunately exhibits significant biases in its outputs. This serves as yet another reminder that strong general performance does not necessarily equate to an unbiased system.

5.2 Human Evaluation

To estimate the accuracy of our gender bias evaluation, we followed again the steps of Stanovsky et al. (2019) [3] and conducted a human validation of the tools. Since our analysis relies on correctly identifying the gender of certain entities in translated sentences, it is important to verify that both the text alignment and the gender extraction tools were performing reliably.

A random sample of 182 sentences from the dataset were selected, representing 5% of the total set of sentences. The translations were selected evenly across all translation models. Each translation was independently annotated by two native Portuguese speakers, who were tasked with identifying the gender of a given entity in the translated sentence. Conducting this evaluation at the sentence level allowed us to account for potential errors in both alignment and gender extraction.

We then compared the gold annotations provided by the humans with the automatic annotations provided by the tools. The agreement between human and automatic annotations was always above 95%. Additionally, the inter-annotator agreement among humans annotators was 98%.

Upon analyzing the cases where human annotations and the tools' output diverged, we found that most tool-related errors involved occupations with identical masculine and feminine forms, distinguished only by the preceding determiners. For example, "the therapist" can be *o terapeuta* (masculine) or *a terapeuta* (feminine).

Some minor discrepancies between the human annotators arose from the translation of certain occupational terms, which were either incorrect or unfamiliar. For example, in European Portuguese, the word "attendant" does not have a single direct translation, it has multiple possible translations that depend on the context. As a result, the translations for this occupation were sometimes slightly incorrect (*empregado*) or included unusual terms (*atendente*), which led to some uncertainty among annotators.

5.3 Translation Quality

We also evaluated translation quality using the COMET-Kiwi metric [36], a reference-free evaluation tool. COMET-Kiwi generates scores ranging from 0 to 1, where 1 indicates high-quality translations, and 0 suggests performance no better than random chance. The results are presented in Table 5.2.

Model	COMET Score(%)
Google Translate	**80.2**
Amazon Translate	78.7
Microsoft Translate	79.5
DeepL	79.8
GPT-4o	79.2
DeepSeek-V3	79.7
Llama-3.2 3B	<u>77.1</u>
Llama-3.2 3B Instruct	78.0
TowerBase 7B	80.0
TowerInstruct 7B	80.0
EuroLLM 9B	80.0
OPUS-MT	79.6
NLLB200 3.3B	79.8
M2M100 1.2B	78.7

Table 5.2: COMET scores for various translation models on the WinoMT dataset.

As illustrated in the table, 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. This suggests that systems with lower bias, such as commercial systems, have successfully incorporated bias mitigation strategies without compromising translation quality. Second, even state of the art translation quality metrics alone are insufficient to capture biases. This aligns with previous concerns about metrics limitations [72] and underscores the need for dedicated bias-focused evaluations to thoroughly assess translation systems.

5.4 Portuguese vs. other Romance Languages

To better understand the performance and bias metrics for Portuguese translations, we conducted a comparative analysis across other Romance languages. Our evaluation included Spanish, French, and Italian.

We started by examining the original study that introduces the WinoMT test suite, conducted in 2019 [3]. After replicating the experiments with updated translations, our findings indicate an improvement over the results reported there for Romance Languages, shown on Table 5.3. 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. Results are shown on Table 5.4. Figure 5.4 provides a visual comparison of Accuracy scores for each language.

	Google Translate			**Microsoft Translate**			**Amazon Translate**		
	Acc	ΔG↓	ΔS↓	**Acc**	ΔG↓	ΔS↓	**Acc**	ΔG↓	ΔS↓
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 5.3: Results for Spanish, French and Italian presented in the 2019 study by Stanovsky et al. [3]

	Google Translate				**Amazon Translate**				**Microsoft Translate**			
	Acc	ΔG↓	ΔS↓	**M:F**↓	**Acc**	ΔG↓	ΔS↓	**M:F**↓	**Acc**	ΔG↓	ΔS↓	**M:F**↓
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

	DeepL				**GPT-4o**				**DeepSeek-V3**			
	Acc	ΔG↓	ΔS↓	**M:F**↓	**Acc**	ΔG↓	ΔS↓	**M:F**↓	**Acc**	ΔG↓	ΔS↓	**M:F**↓
PT	75.4	0.9	21.7	1.47	**59.8**	**9.8**	31.8	**2.10**	47.3	35.2	25.2	5.13
ES	**78.6**	0.9	**17.7**	**1.40**	56.5	11.8	33.9	2.39	**68.2**	3.7	25.7	1.84
FR	69.7	**0.7**	23	1.57	57.9	11.1	**19.3**	2.53	62.4	**0.6**	24.6	**1.55**
IT	54.8	4.8	22.5	1.75	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	ΔG↓	ΔS↓	**M:F**↓	**Acc**	ΔG↓	ΔS↓	**M:F**↓
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	ΔG↓	ΔS↓	**M:F**↓	**Acc**	ΔG↓	ΔS↓	**M:F**↓	**Acc**	ΔG↓	ΔS↓	**M:F**↓
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	ΔG↓	ΔS↓	**M:F**↓	**Acc**	ΔG↓	ΔS↓	**M:F**↓	**Acc**	ΔG↓	ΔS↓	**M:F**↓
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 5.4: Performance of the translation models on the WinoMT dataset across different languages.

Although Spanish, French, and Italian do not support neutral pronouns, we included the full set of 3,888 sentences from the original dataset, neutral examples included, for consistency with the original paper. To ensure a fair comparison, we also computed the Portuguese results using all sentences. As

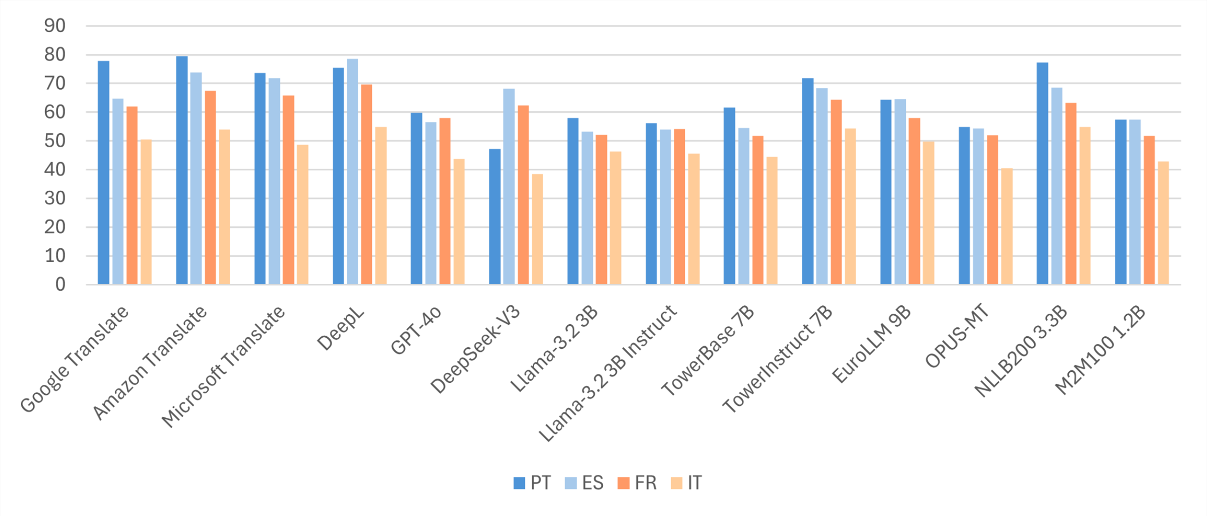


Figure 5.4: Accuracy of each model on the WinoMT dataset, for Portuguese, Spanish, French and Italian.

noted earlier, this slightly skews the results, but since all languages face the same limitation, it remains a reasonable standard for comparison.

Interestingly, the English-Portuguese language pair yielded significantly better results than the other tested languages, across all metrics. In contrast, Italian generally struggled with accuracy and exhibited the highest ΔG, ΔS, and M:F ratio in many of the models. Spanish and French obtained similar results, generally reflecting similar tendencies to Portuguese, though with slightly more biased outcomes.

It is difficult to draw objective conclusions about the underlying reasons for these differences. Unfortunately, this is one of the main challenges when studying biases: they can originate from a wide range of factors. These may include the structure of the language itself, the quantity and nature of training data available for each language, how that data is processed, the architecture of the models, and more. Therefore, it is likely that the observed bias differences stem from an interplay of language features, training data characteristics, and model design.

5.5 Biased Adjectives

Some studies examined how neutral adjectives were translated into masculine and feminine forms [4, 14]. Their findings indicate that gender choices were neither uniform nor evenly distributed, and reveal a strong tendency toward stereotypes. For example, adjectives like “shy” and “desirable” were more frequently translated with female pronouns, while “guilty” and “cruel” were almost exclusively associated with male pronouns.

In their pivotal study, Stanovsky et al. (2019) [3] attempted to “fight bias with bias”. They added the adjectives “handsome” for male entities and “pretty” for female entities to see how this would influence the results. Interestingly, all models showed improved performance once these adjectives were added.

We attempted this approach as well, but unfortunately our results were unreliable. Adding a new word between the determiner and the entity's name ("o belo advogado") led to poorer performance from both the alignment tool and the subsequent gender extraction process. Given that the results were not promising, and considering the time and effort involved to try to fix this issue, we decided not to pursue this hypothesis further and instead prioritize other experiments.

5.6 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]). Given prior research demonstrating that certain sentiment words can impact translation outcomes [13, 14], we were interested in exploring whether the overall sentiment of the sentence might have a similar effect.

To perform sentiment analysis, we utilized Hugging Face's sentiment analysis pipeline with the default DistilBERT model[3]. We then measured the correlation between sentiment (positive/negative) and the gender of translations (masculine/feminine) using Pearson's correlation coefficient.

Pearson's correlation quantifies the strength and direction of a linear relationship between two variables. A coefficient of +1 indicates a perfect positive correlation (when sentiment is positive, the gender tends to be male) , while -1 signifies a perfect negative correlation (when sentiment is positive, the gender tends to be female). A coefficient of 0 implies no correlation. To assess the statistical significance of our findings, we also examined the p-value. A low p-value (typically < 0.05) suggests a meaningful relationship, unlikely to have occurred by chance.

The results of this analysis are presented in Table 5.5.

For most models, the correlation is very close to 0, indicating a negligible relationship between gender and sentiment. However, GPT, Tower Base, and Google Translate yield slightly stronger values, suggesting a minor inverse relationship between gender and sentiment.

Among these models, GPT exhibits the strongest negative correlation, and its very low p-value, approximately zero, indicates statistical significance. Although this correlation is not particularly strong, it does imply a real inverse relationship: translations labeled as "male" tend to be slightly less associated with positive sentiment compared to those labeled as "female". Perhaps unexpectedly, this finding aligns with prior research, which indicated that positive words were often slightly more associated with female entities [13, 14].

Tower Base also shows negative correlations with some degree of statistical significance. However,

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

Model	Correlation	p-value
Google Translate	-0.042	0.014
Amazon Translate	-0.029	0.090
Microsoft Translate	-0.025	0.139
DeepL	-0.034	0.043
GPT-4o	**-0.075**	**0.000**
DeepSeek-V3	-0.023	0.176
Llama-3.2 3B	-0.039	0.021
Llama-3.2 3B Instruct	-0.004	0.817
TowerBase 7B	-0.050	0.003
TowerInstruct 7B	0.006	0.720
EuroLLM 9B	-0.015	0.376
OPUS-MT	-0.037	0.030
NLLB200 3.3B	-0.024	0.153
M2M100 1.2B	-0.022	0.190

Table 5.5: Pearson Correlation and p-values for different models.

these correlations are weaker than GPT's and are likely not substantial enough to indicate a meaningful relationship.

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

Inter-sentence Evaluation

Contents

In this chapter, we explore gender bias in inter-sentence contexts, where gender clues appear in a separate sentence. All experiments and results focus on the inter-2 and inter-3 datasets, two extensions of the WinoMT dataset detailed in Chapter 4.

6.1 Automatic Evaluation

Considering the inter-sentence gender bias, Table 6.1 presents the results of each model on the 500 original sentences, the inter-2 (I2) and inter-3 (I3) datasets. Figure 6.1 visually illustrates the difference in accuracy results between each set of sentences, facilitating comparison.

	Accuracy			$\Delta G\downarrow$			$\Delta S\downarrow$			**M:F**↓		
	Orig.	**I2**	**I3**	**Orig.**	**I2**	**I3**	**Orig.**	**I2**	**I3**	**Orig.**	**I2**	**I3**
Google Translate	83.6	81.2	77.2	2.8	4.0	6.0	22.8	23.7	22.4	1.37	1.50	1.68
Amazon Translate	90.8	50.2	50.2	1.2	45.0	45.1	0.5	22.3	22.3	1.15	8.22	8.42
Microsoft Translate	83.8	49.2	49.2	2.5	38.7	38.1	10.7	22.8	23.2	1.25	5.11	5.02
DeepL	83.2	68.8	48.6	2.7	11.0	38.6	17.4	32.2	27.7	1.34	2.17	5.14
GPT-4o	50.4	50.2	55.0	27.1	33.5	26.5	39.7	37.6	46.0	3.12	4.23	3.40
DeepSeek	51.4	49.8	50.6	36.9	39.5	40.0	29.5	30.8	29.9	5.24	5.54	5.94
Llama-3.2 3B	59.0	56.0	53.0	19.9	21.9	27.9	35.7	32.6	31.7	2.64	2.84	3.43
Llama-3.2 3B Instruct	58.6	63.6	57.4	23.5	18.8	25.0	29.9	20.1	25.4	3.18	2.83	3.37
TowerBase 7B	65.4	70.8	60.0	16.8	11.8	24.3	25.4	23.6	29.1	2.73	2.19	3.42
TowerInstruct 7B	78.8	77.4	71.4	1.7	2.2	3.9	28.1	30.8	28.1	1.18	1.23	1.36
EuroLLM 9B	74.6	75.4	70.6	10.0	8.3	13.2	17.4	13.4	17.4	2.23	1.95	2.52
OPUS-MT	60.4	58.2	57.6	29.1	29.4	32.4	21.4	24.1	19.7	4.86	4.67	5.16
NLLB200 3.3B	80.2	71.6	67.8	6.1	10.0	13.1	22.8	23.2	29.9	1.73	2.13	2.49
M2M100 1.2B	66.8	52.8	49.8	18.3	33.0	36.5	29.0	27.7	29.9	3.11	5.0	5.52

Table 6.1: Results for all metric on the datasets 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.

Analysis of the other metrics reveals that the most significant changes in the inter-sentence datasets occur in ΔG and the M:F ratio, which reach values far higher that those observed in the original dataset. This suggests that, in the absence of sufficient context, many models tend to default to masculine forms.

Among the best performing systems on the original sentences, only Google Translate, EuroLLM,

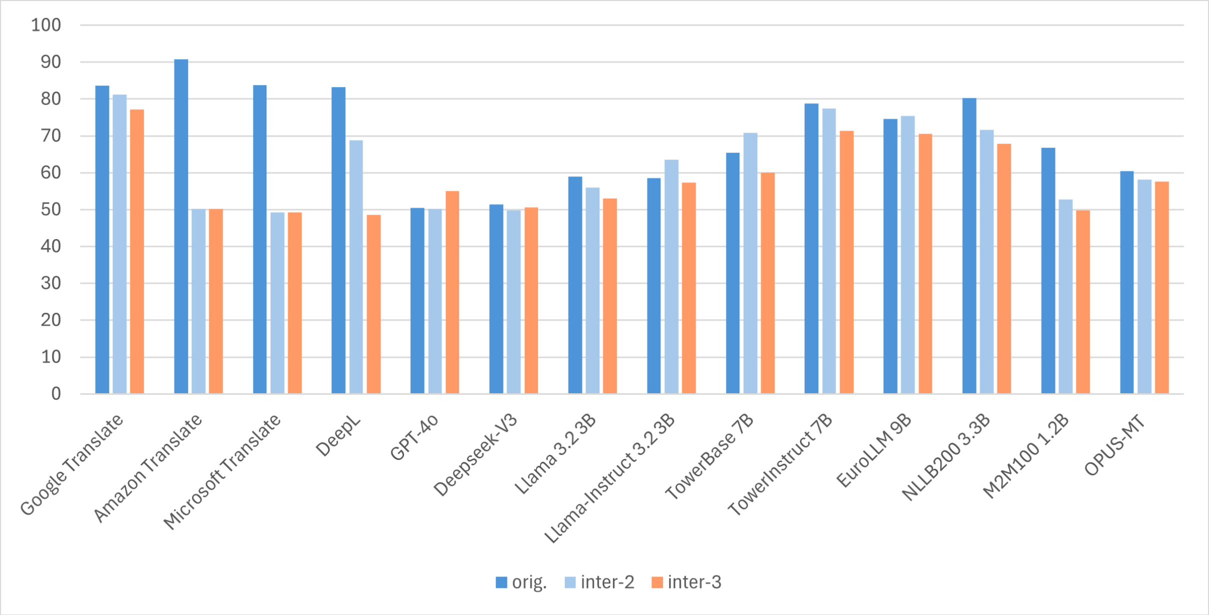


Figure 6.1: Accuracy of each model on the 500 original sentences,the inter-2 and inter-3 datasets.

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.

6.2 Human Evaluation

We also conducted human validation on the inter-sentence datasets to ensure that the alignment and gender extraction tools remained reliable on these sentence structures. For each dataset, 56 instances were randomly sampled, representing slightly over 10% of the total datasets. Once again, the translations were selected evenly across all models. In both cases, agreement between human and automatic annotations exceeded 90%, and inter-annotator agreement was above 95%.

7

Fine-Tuning as a Mitigation Strategy

Contents

Fine-tuning has emerged as a popular mitigation strategy for tackling gender bias. Prior works, such as Saunders and Byrne (2020) [63], Costa Costa-jussá and de Jorge (2020) [66], and Trainotti et al. (2024) [67], have successfully applied this technique. Fine-tuning allows to adapt an already existing model using a smaller, trusted dataset while leveraging existing knowledge. It offers an efficient alternative to other approaches that often involve training models from scratch, a process that is both computationally expensive and technically challenging. In this section, we draw inspiration from the work of Saunders and Byrne (2020) [63], and perform fine-tuning using a tiny, handcrafted, gender-balanced set of sentences.

7.1 Parameter-Efficient Fine-Tuning

Parameter-Efficient Fine-Tuning (PEFT) refers to a set of techniques for fine-tuning models without updating all their weights. Instead, most of the model's weights are frozen, and only a small number of trainable modules are modified. By preserving most of the model's pre-existing parameters, PEFT helps maintain generalization ability while still allowing adaptation to new tasks.

One major challenge when fine-tuning on small datasets is *catastrophic forgetting* [63]. This phenomenon occurs when a model adapts too strongly to the new data, resulting in degraded performance on tasks that it previously handled well, such as general translation. This makes PEFT particularly well-suited to our experiments.

Additionally, since only the trainable parameters need to be stored in GPU, PEFT significantly reduces computational cost and memory usage. This allows to fine-tune larger models even under hardware constraints.

In our work, we used Low-Rank Adaptation (LoRA) [73], a popular PEFT method. LoRA works by adding smaller matrices (low-rank matrices) to the attention layers. Only these new parameters are optimized, which drastically reduces the number of trainable parameters.

7.2 Experimental setup

In this section, we describe the experimental setup used to evaluate our fine-tuning approach.

7.2.1 Model

Due to time and hardware constraints, we selected the NLLB200-1.3B model for fine-tuning. We chose the NLLB family because NLLB 3.3B had previously shown a good balance between good performance and small size. It achieved strong results in our earlier bias assessment experiments while remaining significantly smaller than other models with comparable performance. However, the 3.3B checkpoint proved to be too computationally demanding for our available resources. Therefore, we opted for the next smaller version, NLLB-1.3B. While the use of a smaller model slightly reduces the direct relevance of our findings, we believe that the results obtained still provide valuable insights into the effectiveness of fine-tuning for gender bias mitigation.

7.2.2 Dataset

As for the dataset used for fine-tuning, we use the dataset provided by Saunders and Bryne [63]. There are two versions of this dataset: the original version, and the no-overlap version, which excludes occupations present in the WinoMT dataset. In the original dataset, all sentences follow the structure:

"The <profession> finished his/her job."

Example 7: Original dataset

Fem: *The professor finished her work.*

Masc: *The professor finished his work.*

This dataset contains 388 sentences, with each occupation appearing once in the masculine form and once in the feminine form. This exposes the model to both gendered expressions.

In the no-overlap dataset, many occupations are excluded, resulting in only 216 sentences. To increase the number back to 388, additional adjectives are used to create new sentences. The structure of these new sentences are as follows:

"The <adjective> man/woman finished his/her job".

Example 8: No-overalp dataset

Fem: *The tall woman finished her work.*

Masc: *The tall man finished his work.*

To create the Portuguese references for the dataset, we did a first-pass using a translation motor (Microsoft Translate). Then, we manually reviewed each translation and corrected any errors or unnatural word choices.

7.2.3 Hyperparameters

The fine-tuning was performed using the Hugging Face's Seq2SeqTrainer class. We used the Optuna library[1] [74] to search for the most efficient hyperparameters. Optuna is a Python library used for automated hyperperamitization optimization. We conducted 8 rounds for each dataset, exploring learning rate, number of epochs, and weight decay. We also tuned LoRA-specific hyperparameters, including the dimension of low-rank matrices (r), the scaling factor for the matrices (alpha), and dropout. The number of rounds was limited to 8, as additional rounds exceeded our available computational resources.

The datasets were split into a 90/10 train-validation set, and BLEU scores, computed with SacreBLEU, were used as the evaluation metric.

For the sake of reproducibility, we report the best hyperparameters found for each model, which were subsequently used for fine-tuning.

In the original dataset, the best trial achieved a BLEU of 87.1 in the validation set, with the following hyperparameters:

- **Learning rate:** 5.6e-5
- **Epochs:** 3
- **Weight decay:** 0.09
- **r:** 8
- **Alpha:** 32
- **Dropout:** 0.1

In the no-overlap dataset, the best trial obtained a BLEU of 91.3 in the validation set, with the following hyperparameters:

- **Learning rate:** 2.5e-4
- **Epochs:** 10
- **Weight decay:** 0.04
- **r:** 6
- **Alpha:** 32
- **Dropout:** 0.09

All other parameters were kept at their default values, as defined by the respective HuggingFace implementations.

[1] https://optuna.org/

7.3 Results and Discussion

Table 7.1 shows the bias results for the fine-tuned models on the WinoMT dataset. The baseline for comparison is the original NLLB200-1.3B model, without any fine-tuning.

Model	**Acc**	$\Delta G\downarrow$	$\Delta S\downarrow$	**M:F**$\downarrow$
Baseline	77.3	5.8	27.3	1.67
Fine-tuned (original)	79.2	4.6	26	1.56
Fine-tuned (no-overlap)	**90.7**	**0.8**	**8.9**	**1.15**

Table 7.1: Performance of the fine-tuned models on WinoMT.

The results show that fine-tuning significantly reduces gender bias, particularly when using the no-overlap dataset. It was somewhat unexpected that the model fine-tuned on the no-overlap dataset would outperform the one fine-tuned on the original dataset. This may be attributed to more effective hyper-parameter optimization, as suggested by the higher BLEU scores obtained through Optuna. However, these results on the no-overlap dataset also provide an important takeaway: the model is not simply memorizing the gendered vocabulary. Despite the lack of shared occupational terms with the WinoMT dataset, it shows substantial improvements on this test set. This suggests that the model is learning to generalize and diversify gender usage, which is crucial for meaningful bias mitigation.

We observe that not only does ΔG decrease substantially, reaching near-perfect scores in the no-overlap fine-tuning, but ΔS also improves notably for this version.

We also note that the model fine-tuned on the no-overlap dataset outperforms all other models tested in Chapter 5 across all metrics. The model trained on the original dataset also achieves strong results compared to the other models, though its performance remains very similar to that of the original NLLB200-3.3B.

To ensure we are not falling victims to catastrophic forgetting, and gender-focused fine-tuning does not degrade the overall performance of our fine-tuned model, we evaluate the general translation quality of our models using a separate dataset covering a diverse range of sentences and scenarios.

For this purpose, we chose the WMT24++ dataset [61], and assess the quality of the translations using both BLEU and COMET scores. This dataset is particularly challenging, consisting of 998 entries that span literary texts, news articles, and even social media posts. Table 7.2 show the results of each model in the general translation task.

Model	BLEU	COMET(%)
Baseline	25.6	23.9
Fine-tuned (original)	26.0	**25.7**
Fine-tuned (no-overlap)	**26.6**	23.9

Table 7.2: Performance of fine-tuned models on the WMT24++ dataset.

We note that although the baseline scores were not exceptionally high, they did not deteriorate after fine-tuning. In fact, they improved slightly. This suggests that we successfully prevented catastrophic forgetting during the fine-tuning process.

While our analysis was limited in scope, the results clearly demonstrate the potential of fine-tuning as a practical bias mitigation strategy. With relatively low computational cost and simple implementation, fine-tuning was able to significantly reduce bias while maintaining general translation performance.

8

Conclusion

Contents

In this study, we explored the presence of gender bias in English-to-Portuguese translation. We found 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.

We also compared the results for Portuguese with those of other Romance languages, revealing that 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.

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.

Finally, we explored the potential of fine-tuning a pre-existing model on a small, handcrafted, gender-balanced dataset as a strategy for mitigating gender bias in the machine translation to Portuguese. Our approach demonstrated that fine-tuning can effectively reduce gender bias in translation outputs, achieving significant improvements across multiple bias metrics. Importantly, these improvements were achieved without compromising the overall translation performance.

8.1 Limitations

8.1.1 Data

In this work, our analysis was restricted to bias related to occupational stereotypes, which is only one of many manifestations of gender bias. The structured nature of the sentences in our dataset, designed to isolate occupational stereotypes, may also lead models to learn these specific patterns without addressing broader issues of bias.

Moreover, this study only considers binary gender forms (female and male), overlooking non-binary and gender-neutral representations.

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.

8.1.2 Models

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 our evaluation. For the fine-tuning experiments, we selected the NLLB-1.3B model because of its manageable computational requirements. However, given this model's relatively small size and prominence, it may not be the most relevant model to explore. Nevertheless, we hypothesize that similar fine-tuning approaches could yield strong results for larger models as well.

Additionally, 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 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.

8.2 Future Work

While this thesis provides a comprehensive analysis of gender bias in current translation models, several areas remain open for further research.

First, this work focuses on detecting and evaluating the extent of gender bias in current translation models for the English-Portuguese language pair. It lays the foundation for future research on specific strategies to mitigate these biases. In Chapter 7 we explore fine-tuning on a small dataset, which demonstrated encouraging results, but Chapter 3 discusses various other mitigation strategies we also consider feasible for this language pair.

Additionally, our study primarily focuses on gender bias, particularly in occupational stereotypes. Expanding the scope to other forms of biases, such as racial bias or age bias, would contribute to more fair and equitable machine translation.

Another promising avenue for exploration is gender-neutral translation. Developing tools and evaluation methods for neutral translations would be a significant step toward gender-fair language and translation practices.

We hope this research encourages future exploration into detecting and mitigating bias in AI-driven language technologies.

Bibliography

[1] J. Alammar. (2018) Visualizing a neural machine translation model (mechanics of seq2seq models with attention). Accessed 08/05/2024. [Online]. Available: https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

[2] ——. (2018) The illustrated transformer. Accessed 08/05/2024. [Online]. Available: https://jalammar.github.io/illustrated-transformer/

[3] G. Stanovsky, N. A. Smith, and L. Zettlemoyer, "Evaluating gender bias in machine translation," *arXiv preprint arXiv:1906.00591*, 2019.

[4] M. O. R. Prates, P. H. C. Avelar, and L. Lamb, "Assessing gender bias in machine translation – a case study with google translate," no. arXiv:1809.02208, 2019, arXiv:1809.02208 [cs]. [Online]. Available: http://arxiv.org/abs/1809.02208

[5] S. Ghosh and A. Caliskan, "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*, 2023, pp. 901–912.

[6] B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, and M. Turchi, "Gender bias in machine translation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 845–874, 2021.

[7] B. Savoldi, S. Papi, M. Negri, A. Guerberof-Arenas, and L. Bentivogli, "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*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 18 048–18 076. [Online]. Available: https://aclanthology.org/2024.emnlp-main.1002/

[8] H. Gonen and K. Webster, "Automatically identifying gender issues in machine translation using perturbations," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1991–1995. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.180

[9] E. Vanmassenhove, “Gender bias in machine translation and the era of large language models,” *arXiv preprint arXiv:2401.10016*, 2024.

[10] G. Attanasio, F. M. Plaza del Arco, D. Nozza, and A. Lauscher, “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*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 3996–4014. [Online]. Available: https://aclanthology.org/2023.emnlp-main.243/

[11] J. Zhao, Y. Ding, C. Jia, Y. Wang, and Z. Qian, “Gender bias in large language models across multiple languages,” *arXiv preprint arXiv:2403.00277*, 2024.

[12] M. Menezes, M. A. Farajian, H. Moniz, and J. V. Graça, “A context-aware annotation framework for customer support live chat machine translation,” in *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, M. Utiyama and R. Wang, Eds. Macau SAR, China: Asia-Pacific Association for Machine Translation, Sep. 2023, pp. 286–297. [Online]. Available: https://aclanthology.org/2023.mtsummit-research.24/

[13] W. I. Cho, J. W. Kim, S. M. Kim, and N. S. Kim, “On measuring gender bias in translation of gender-neutral pronouns,” in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 173–181. [Online]. Available: https://aclanthology.org/W19-3824

[14] W. I. Cho, J. Kim, J. Yang, and N. S. Kim, “Towards cross-lingual generalization of translation gender bias,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 449–457. [Online]. Available: https://doi.org/10.1145/3442188.3445907

[15] B. Friedman and H. Nissenbaum, “Bias in computer systems,” *ACM Trans. Inf. Syst.*, vol. 14, no. 3, p. 330–347, jul 1996. [Online]. Available: https://doi.org/10.1145/230538.230561

[16] K. Crawford, “The trouble with bias,” 2017, in Conference on Neural Information Processing Systems (NIPS) – Keynote, Long Beach, USA.

[17] P. Koehn, “Europarl: A parallel corpus for statistical machine translation,” in *Proceedings of Machine Translation Summit X: Papers*, Phuket, Thailand, Sep. 13-15 2005, pp. 79–86. [Online]. Available: https://aclanthology.org/2005.mtsummit-papers.11

[18] E. Vanmassenhove, C. Hardmeier, and A. Way, “Getting gender right in neural machine translation,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3003–3008. [Online]. Available: https://aclanthology.org/D18-1334

[19] L. Bentivogli, "Good but not always fair, tackling gender bias in automatic translation," 2022, in Translating and the Computer - TC44 – Keynote, Luxembourg.

[20] M. R. Costa-jussà, C. Escolano, C. Basta, J. Ferrando, R. Batlle, and K. Kharitonova, "Gender bias in multilingual neural machine translation: The architecture matters," *arXiv preprint arXiv:2012.13176*, 2020.

[21] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in machine translation," *Engineering*, vol. 18, pp. 143–153, 2022.

[22] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[23] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[26] D. Stahlberg, F. Braun, L. Irmen, and S. Sczesny, "Representation of the sexes in language," *Social Communication*, pp. 163–187, 01 2007.

[27] A. Stafanovičs, T. Bergmanis, and M. Pinnis, "Mitigating gender bias in machine translation with target gender annotations," in *Proceedings of the Fifth Conference on Machine Translation*, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, Y. Graham, P. Guzman, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, and M. Negri, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 629–638. [Online]. Available: https://aclanthology.org/2020.wmt-1.73

[28] D. Saunders, R. Sallis, and B. Byrne, "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, M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, Eds. Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 35–43. [Online]. Available: https://aclanthology.org/2020.gebnlp-1.4

[29] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[30] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[31] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "Comet: A neural framework for mt evaluation," *arXiv preprint arXiv:2009.09025*, 2020.

[32] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "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*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, Aug. 8-12 2006, pp. 223–231. [Online]. Available: https://aclanthology.org/2006.amta-papers.25

[33] M. Elaraby, A. Y. Tawfik, M. Khaled, H. Hassan, and A. Osama, "Gender aware spoken language translation applied to english-arabic," in *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*. IEEE, 2018, pp. 1–6.

[34] J. Escudé Font and M. R. Costa-jussà, "Equalizing gender bias in neural machine translation with word embeddings techniques," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 147–154. [Online]. Available: https://aclanthology.org/W19-3821

[35] A. Piergentili, B. Savoldi, D. Fucci, M. Negri, and L. Bentivogli, "Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 14 124–14 140. [Online]. Available: https://aclanthology.org/2023.emnlp-main.873

[36] R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. C. de Souza, T. Glushkova, D. Alves, L. Coheur, A. Lavie, and A. F. T. Martins, "CometKiwi: IST-unbabel 2022

submission for the quality estimation shared task," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, and M. Zampieri, Eds. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 634–645. [Online]. Available: https://aclanthology.org/2022.wmt-1.60/

[37] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "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*, 2020.

[38] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020.

[39] M. Straka, J. Hajič, and J. Straková, "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)*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 4290–4297. [Online]. Available: https://aclanthology.org/L16-1680

[40] N. Mamede, "String: An hybrid statistical and rule-based natural language processing chain for portuguese."

[41] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[42] M.-C. De Marneffe, C. D. Manning, J. Nivre, and D. Zeman, "Universal dependencies," *Computational linguistics*, vol. 47, no. 2, pp. 255–308, 2021.

[43] M. Gonçalves, L. Coheur, J. Baptista, and A. Mineiro, "Avaliação de recursos computacionais para o português," *Linguamática*, vol. 12, no. 2, pp. 51–68, 2021.

[44] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[45] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[46] T. Kocmi, E. Avramidis, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, M. Karpinska, P. Koehn, B. Marie, C. Monz, K. Murray,

M. Nagata, M. Popel, M. Popović, M. Shmatova, S. Steingrímsson, and V. Zouhar, "Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet," in *Proceedings of the Ninth Conference on Machine Translation*, B. Haddow, T. Kocmi, P. Koehn, and C. Monz, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 1–46. [Online]. Available: https://aclanthology.org/2024.wmt-1.1/

[47] D. M. Alves, J. Pombal, N. M. Guerreiro, P. H. Martins, J. Alves, A. Farajian, B. Peters, R. Rei, P. Fernandes, S. Agrawal, P. Colombo, J. G. C. de Souza, and A. F. T. Martins, "Tower: An open multilingual large language model for translation-related tasks," 2024. [Online]. Available: https://arxiv.org/abs/2402.17733

[48] P. H. Martins, P. Fernandes, J. Alves, N. M. Guerreiro, R. Rei, D. M. Alves, J. Pombal, A. Farajian, M. Faysse, M. Klimaszewski *et al.*, "Eurollm: Multilingual language models for europe," *arXiv preprint arXiv:2409.16235*, 2024.

[49] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[50] J. Tiedemann and S. Thottingal, "OPUS-MT – building open translation services for the world," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, and M. L. Forcada, Eds. Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 479–480. [Online]. Available: https://aclanthology.org/2020.eamt-1.61

[51] J. Tiedemann, "Parallel data, tools and interfaces in opus." in *Lrec*, vol. 2012. Citeseer, 2012, pp. 2214–2218.

[52] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary *et al.*, "Beyond english-centric multilingual machine translation," *Journal of Machine Learning Research*, vol. 22, no. 107, pp. 1–48, 2021.

[53] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang, "Mitigating gender bias in natural language processing: Literature review," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1630–1640. [Online]. Available: https://aclanthology.org/P19-1159

[54] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "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), M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 15–20. [Online]. Available: https://aclanthology.org/N18-2003

[55] R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme, “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)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 8–14. [Online]. Available: https://aclanthology.org/N18-2002

[56] L. Bentivogli, B. Savoldi, M. Negri, M. A. Di Gangi, R. Cattoni, and M. Turchi, “Gender in danger? evaluating speech translation technology on the MuST-SHE corpus,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 6923–6933. [Online]. Available: https://aclanthology.org/2020.acl-main.619

[57] B. Savoldi, M. Gaido, M. Negri, and L. Bentivogli, “Test suites task: Evaluation of gender fairness in mt with must-she and ines,” *arXiv preprint arXiv:2310.19345*, 2023.

[58] A. Currey, M. Nădejde, R. Pappagari, M. Mayer, S. Lauly, X. Niu, B. Hsu, and G. Dinu, “Mt-geneval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation,” *arXiv preprint arXiv:2211.01355*, 2022.

[59] P. Lison and J. Tiedemann, “Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles,” 2016.

[60] K. Wołk and K. Marasek, “Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs,” *Procedia Technology*, vol. 18, pp. 126–132, 2014.

[61] D. Deutsch, E. Briakou, I. Caswell, M. Finkelstein, R. Galor, J. Juraska, G. Kovacs, A. Lui, R. Rei, J. Riesa *et al.*, “Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects,” *arXiv preprint arXiv:2502.12404*, 2025.

[62] C. R. S. Basta, M. Ruiz Costa-Jussà, and J. A. Rodríguez Fonollosa, “Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information,” in *Proceedings of the The Fourth Widening Natural Language Processing Workshop*. Association for Computational Linguistics, 2020, pp. 99–102.

[63] D. Saunders and B. Byrne, “Reducing gender bias in neural machine translation as a domain adaptation problem,” *arXiv preprint arXiv:2004.04498*, 2020.

[64] L. Veloso, L. Coheur, and R. Ribeiro, “A rewriting approach for gender inclusivity in Portuguese,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8747–8759. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.585

[65] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” 2016.

[66] M. R. Costa-jussà and A. de Jorge, “Fine-tuning neural machine translation on gender-balanced datasets,” in *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, Eds. Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 26–34. [Online]. Available: https://aclanthology.org/2020.gebnlp-1.3/

[67] R. Trainotti Rabonato, E. Milios, and L. Berton, “Gender-neutral english to portuguese machine translator: Promoting inclusive language,” in *Brazilian Conference on Intelligent Systems*. Springer, 2024, pp. 180–195.

[68] C. Dyer, V. Chahuneau, and N. A. Smith, “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*, 2013, pp. 644–648.

[69] J. Graça, J. P. Pardal, L. Coheur, and D. Caseiro, “Building a golden collection of parallel multi-language word alignments,” in *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, ser. LREC. Marrakech, Morocco: LREC 2008, May 28-30 2008.

[70] F. J. Och and H. Ney, “Improved statistical alignment models,” in *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Hong Kong: Association for Computational Linguistics, Oct. 2000, pp. 440–447. [Online]. Available: https://aclanthology.org/P00-1056/

[71] A. Rademaker, F. Chalub, L. Real, C. Freitas, E. Bick, and V. de Paiva, “Universal dependencies for portuguese,” in *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, Pisa, Italy, September 2017, pp. 197–206. [Online]. Available: http://aclweb.org/anthology/W17-6523

[72] C. Zerva, F. Blain, J. G. C. De Souza, D. Kanojia, S. Deoghare, N. M. Guerreiro, G. Attanasio, R. Rei, C. Orasan, M. Negri, M. Turchi, R. Chatterjee, P. Bhattacharyya, M. Freitag, and A. Martins, "Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE?" in *Proceedings of the Ninth Conference on Machine Translation*, B. Haddow, T. Kocmi, P. Koehn, and C. Monz, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 82–109. [Online]. Available: https://aclanthology.org/2024.wmt-1.3/

[73] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[74] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.