Gender Bias Evaluation in English-Portuguese Automated Translation Outputs

Xiaolan Xu
Faculty of Arts and Humanities, University of Lisbon, Portugal

Sara Mendes
Center of Linguistics of the University of Lisbon and Faculty of Arts and Humanities, University of Lisbon, Portugal

Abstract

Automated translation (AT) plays a pivotal role in breaking down language barriers and fostering cross-cultural communication. However, gender bias in AT remains a pressing concern, particularly for language pairs where the source language generally does not have grammatical gender and the target language does. In this context, the present research aims to assess gender bias in English-Portuguese AT outputs. Two machine translation (MT) systems and two LLMs are evaluated in this study, examining the correlation between translated gender and societal stereotypes, and how the different models handle nouns whose gender is unknown. The outputs of two commercial MT systems (Google Translate and DeepL Translator) and two general-purpose LLMs (ChatGPT-4o and DeepSeek-V3), translating from English into European Portuguese, were analyzed. We used a subset of the WinoMT challenge set to assess gender rendering accuracy across three conditions: gender-defined nouns, gender-undefined nouns, and anaphoric pronoun resolution. Our main goal is therefore to evaluate the accuracy of gender information transmission and to understand the extent to which gender bias is present in AT outputs. Our results indicate that all four models struggle to faithfully convey gender information from source to target, with male-centric outputs predominating across conditions and MT systems showing a more pronounced bias. This work contributes a preliminary cross-system comparison for an underexplored language pair and lays the groundwork for larger-scale evaluations of gender-inclusive AT.

2012 ACM Subject Classification Computing methodologies → Machine translation; Social and professional topics → Gender

Keywords and phrases Gender Bias, Large Language Models (LLMs), Neural Machine Translation (NMT), English-Portuguese Translation

Digital Object Identifier 10.4230/OASIcs.SLATE.2026.8

Supplementary Material *Dataset*: `https://github.com/xu-xiaolan/gender-bias-en-pt-at` [20] archived at `swh:1:dir:f3d848b15460545b49033fd09dfa2af692d74510`

Funding We acknowledge the support of Fundação para a Ciência e a Tecnologia (UID/00214: Centro de Linguística da Universidade de Lisboa).

1 Introduction

In our globally interconnected world, machine translation (MT) plays a pivotal role in breaking down language barriers and fostering cross-cultural communication. After decades of development, since the concept of MT was created, the current state-of-the-art MT system is neural machine translation (NMT). NMT uses deep neural networks to model the translation process end-to-end. Unlike earlier models like rule-based machine translation (RBMT) or statistical machine translation (SMT), which relied on separate components for tasks such as language modeling and phrase alignment, NMT learns to translate directly from source to target language using large datasets. This results in more fluent and contextually accurate translations. Nowadays, with the development of technology, Large Language Models (LLMs) perform very well in text processing, including translating text.

As highly data-driven approaches, NMT and LLMs rely heavily on extensive datasets and corpora for effective training and learning. However, this reliance on data introduces a significant challenge: the inherent biases present in natural language data. For instance, gender bias often permeates training datasets, with a prevalence of male-centric examples over female-centric ones. Consequently, systems trained on such data may inadvertently perpetuate or even amplify these biases [21].

Gender bias can be manifested in various ways, from the use of gendered pronouns to the translation of gender-specific terms, resulting in errors and reducing output quality, especially when the target language has grammatical gender, like Portuguese. In languages that have grammatical gender distinctions, automated translation (AT) systems may assign masculine pronouns by default or prioritize male-centric interpretations of ambiguous expressions, e.g., translating secretaries and nurses as females, and programmers and doctors as males, regardless of context. This distorts the intended meaning of the original text, leading to incorrect translations. Therefore, addressing gender bias in AT systems is not just a matter of linguistic accuracy but also a crucial step toward improving the performance of automated translation.

Gender bias in machine translation has attracted a lot of attention. Stanovsky et al. [18] present the first challenge set and evaluation protocol to measure the effect of gender bias in MT, and this research reveals the obvious existence of gender-biased translation inaccuracies in four widely used industrial MT systems and two contemporary academic MT. Other researchers have outlined different strategies to reduce gender bias in MT, such as using word embeddings [6] and transfer learning [15], or creating a new learning framework [5]. These studies involve many languages, such as English, German, Spanish, Italian, Finnish, Turkish, Korean, etc., and the translation language pairs mainly involve English. Meanwhile Portuguese has not attracted much attention, while gender expressions in English and Portuguese are highly asymmetrical: in English, there is only lexical gender (*mother/father*), and pronominal gender (*she/he/it*); while in Portuguese, nouns have morphological endings to express the gender class as masculine or feminine (*amigo/amiga*), and due to morphosyntactic agreement, several parts of speech besides the noun carry gender inflections (e.g., determiners, adjectives).

In this context, the present research aims to assess gender bias in English-Portuguese AT outputs. It assesses gender bias in the outputs produced by two commercial NMT systems (i.e., Google Translate and DeepL) and two general-purpose LLMs (i.e., ChatGPT-4o and DeepSeek-V3), with the objectives of evaluating the accuracy of gender information transmission in English-Portuguese AT, analyzing the correlation between gender translation and societal stereotypes, and examining how gender-neutral or unknown gender phrases are handled in AT.

This research is designed to answer three research questions:

- **(i)** Do selected systems correctly convey the gender information of the source text in English-Portuguese AT?
- **(ii)** Does the translation output reproduce or amplify gender stereotypes?
- **(iii)** How do the systems tested handle the unknown or neutral gender from the source to the target?

These research questions will allow us to assess how contemporary automated translation systems handle gender representation. The remainder of this work is structured as follows: Section 2 reviews prior work on gender bias in MT and outlines the gender relevant grammatical features of English and Portuguese. Section 3 details our methodological approach, including data selection and annotation aspects. Section 4 presents the data analysis and findings, and section 5 offers the conclusion.

2 Background

The theoretical background for this study necessarily includes existing research on gender bias in automated translation (Section 2.1), as well as the linguistic description of grammatical gender systems of English and Portuguese (Section 2.2).

2.1 Gender bias in AT

Gender bias is a significant issue in AT, leading to ongoing research efforts in developing bias mitigation techniques. This phenomenon may lead to misgendered translations, resulting, among other harms, in the perpetuation of stereotypes and prejudices.

Gender biases in AT can be manifested by, for example, producing disproportionately fewer feminine forms or consistently translating professional titles in a gender-stereotypical manner [1]. These biases lead to inaccuracies in AT outputs, requiring the post-editing of AT outputs to correctly reflect feminine gender. This results in greater time and effort, which in turn leads to higher financial costs [16].

Stanovsky et al. [18] present the first challenge set, called WinoMT, and the corresponding evaluation protocol to measure the effect of gender bias in MT. This research reveals that four widely used industrial MT systems and two contemporary state-of-the-art academic MT models exhibit notable gender-biased translation inaccuracies across all the examined target languages [18].

Although a systematic study of such biases can be difficult, Prates et al. [11] believe that MT tools can be exploited through gender neutral languages to yield a window into the phenomenon of gender bias in AI. Prates et al. [11] use it to build sentences in constructions like "He/She is an Engineer" (where "Engineer" is replaced by different occupation target nouns of interest) in 12 different gender-neutral languages such as Hungarian, Chinese, Yoruba, and several others. Font et al. [6] take advantage of the fact that word embeddings are used in NMT to propose a method to equalize gender biases in neural machine translation using these representations. They experiment and analyze the integration of two debiasing techniques over GloVe embeddings in the Transformer translation architecture. Cho et al. [2] propose a scheme for making up a test set that evaluates gender bias in a MT system working with Korean, a language with gender-neutral pronouns.

Saunders and Byrne [15] use transfer learning on a small set of trusted, gender-balanced examples to reduce gender bias, and solve the "catastrophic forgetting" problem in adaptation and in inference. Fleisig et al. [5] present a learning framework that alleviate gender bias in seq2seq machine translation. Costa-jussà et al. [4] discuss from an algorithmic perspective whether the chosen architecture, when trained with the same data, influences the level of gender bias, and they propose two methods for interpreting and studying gender bias in MT based on source embeddings and attention.

Recent work has begun to examine gender bias in LLMs, revealing that base LLMs exhibit a higher degree of gender bias than NMT models in English-to-Catalan and English-to-Spanish translation [14]. Similarly, Popović and Lapshinova-Koltunski [10] show that ChatGPT introduces different gender bias patterns when compared to MT systems. These findings suggest that model architecture plays a meaningful role in bias behavior, motivating cross-system comparative analyses.

Table 1 English pronouns [19, p. 13].

			Personal pronouns		Possessive pronouns		Reflexive pronouns
			subjective case	objective case	determiner function	nominal function	
1p	sg.		I	me	my	mine	myself
	pl.		we	us	our	ours	ourselves
2p	sg.		you		your	yours	yourself
	pl.						yourselves
3p	sg.	masc.	he	him	his		himself
		fem.	she	her		hers	herself
		non-personal	it		its		itself
	pl.		they	them	their	theirs	themselves

2.2 Grammatical gender in English and Portuguese

Gender is defined as "classes of nouns reflected in the behavior of associated words [7, p. 231]". In other words, when a language has, for instance, three genders, it means that the nouns in that language can be divided into three distinct classes [3, p. 4]. These classes are syntactically identifiable through the agreement patterns expressed within the nominal phrase (determiners, adjectives), outside the nominal phrase (pronouns, verbs, predicates) and partially also through the noun triggering the agreement itself [17, p. 8]. Therefore, gender functions as a morphosyntactic feature of language since it is required for agreement [8].

In English, there are three genders: masculine, feminine and neuter. These three genders are categorized based on semantics, the criteria being humanness and the sex of the relevant referents, for example, *boy*, *girl*, *dog*, *stone* and *love*. This type of gender assignment system is a semantic gender assignment system [8]. The only domain of grammar where gender is visible in English, i.e. where agreement is triggered, is the pronominal system [17]. As shown in Table 1, it is obvious that only the singular 3rd person has three different forms for different genders, in which masculine pronouns are used for male human referents (*he/him/his/himself*), feminine pronouns for human female referents (*she/her/hers/herself*) while neuter is used for all other entities (*it/its/itself*)[1]. In their anaphoric use, these pronouns replace the relevant nouns.

It is also important to note that in English, there is a group of lexical items that can refer to either a male or female representative of the human species, for example, developer, hairdresser, driver, housekeeper, farmer, etc. These nouns are classified as "personal dual gender" [12].

Siemund [17, p. 161] argues that for these nouns, the choice of pronouns is determined by the referent rather than the noun itself, and that this group includes many profession-related terms.

Since the choice of pronouns depends on the referent rather than on the nouns themselves, pronouns in co-reference relations with such nouns can, in turn, reveal the gender of the referent associated with them. This study focuses precisely on such lexical items.

Conversely, Portuguese has a grammatical gender system with two categories: masculine and feminine. These two genders are "semantically consistent in that nouns referring clearly to males are masculine ($aluno/gato_{[\text{masc.}]}$), those referring clearly to females are feminine

[1] The use of pronouns discussed here is limited to general rules. In actual language use, there are exceptions. For instance, animals, though not human, may be referred to as "he" or "she" based on their biological sex. Similarly, in the case of inanimate objects like "ship", the feminine pronoun "she" is often used. These usages are influenced by emotional and cultural factors, which will not be explored in detail here.

(*aluna*/*gata*$_{[\text{fem.}]}$); but the gender of other nouns is for the most part arbitrary (*copo*$_{[\text{masc.}]}$, *janela*$_{[\text{fem.}]}$) [7, p. 232]". This type of system is referred to as a semantic-and-formal gender assignment system, where the assignment is based on both meaning and form – specifically phonological and morphological features, or a combination of both [8].

Gender is an inherent property of nouns, and, as mentioned earlier, gender agreement occurs both within and beyond the nominal phrase. In English, gender agreement is limited to the pronominal system. In contrast, Portuguese exhibits a broader agreement domain, including pronouns (see (1a)), determiners (1b), adjectives (1c), numerals (1d), and verb participles (1e) [13].

(1) **(a)** *A Ana viu um gato bonito e ela gostou dele.* (Ana saw a$_{[\text{masc.}]}$ beautiful$_{[\text{masc.}]}$ cat$_{[\text{masc.}]}$ and she liked it$_{[\text{masc.}]}$.)

(b) masculine determiners: *o gato* (the cat), *os gatos* (the cats)
feminine determiners: *a gata* (the cat), *as gatas* (the cats)

(c) masculine adjectives: *um trabalho duro* (hard work)
feminine adjectives: *uma tarefa dura* (hard task)

(d) common cardinal numerals in masculine: *vinte e dois*/*duzentos ovos* (twenty-two/two hundred eggs)
common cardinal numerals in feminine: *vinte e duas*/*duzentas bananas* (twenty-two/two hundred bananas)

(e) verbal passive clauses: *O portão foi aberto*/*A janela foi aberta pelo gato.* (The gate$_{[\text{masc.}]}$ was opened$_{[\text{masc.}]}$/The window$_{[\text{fem.}]}$ was opened$_{[\text{fem.}]}$ by the cat.)

To sum up, English and Portuguese differ significantly in how they express gender. The linguistic structural differences introduce challenges in maintaining accurate and unbiased gender representation during AT, especially when translating from a less gendered language (English) into a highly gendered one (Portuguese).

3 Methodology

This study involved three main stages: translation, annotation, and data analysis, incorporating both quantitative and qualitative methods. First, we translated 120 original sentences from the WinoMT dataset [18], using each of the four selected models, applying the same input segments to ensure consistency in the comparison (see section 3.1). Translations were performed sentence by sentence, as the WinoMT dataset consists of isolated sentences specifically designed to test local contextual interpretation – particularly of pronouns and semantic roles. In this study, we focus on the role of pronouns in the generation of accurate gendered translations in European Portuguese.

As aforementioned, the four models considered in this work include two commercial NMT systems (Google Translate[2] and DeepL Translator[3]) and two LLMs (ChatGPT-4o[4] and DeepSeek-V3[5]). These systems were selected based on their popularity, demonstrated translation quality, technological reliability, and free public accessibility. To ensure the uniformity, both MT systems were set to output translation in European Portuguese, and both LLMs received the same prompt: "*Please translate this sentence into European Portuguese.*" This controlled for the variant of the language output produced and ensured comparability

[2] https://translate.google.com/
[3] https://www.deepl.com/translator
[4] https://chatgpt.com/
[5] https://chat.deepseek.com/

across systems. The overall goal of this methodology is to assess how each system handles gender-sensitive structures during translation, particularly with regard to grammatical gender in European Portuguese.

The four systems serve as the independent variable, allowing for a systematic comparison of their outputs under controlled conditions. The dependent variables are the features annotated for each translation output: the grammatical gender assigned to gender-defined and gender-undefined occupation nouns, the stereotypical or anti-stereotypical nature of those gender assignments, and the recoverability of gender information encoded in source pronouns in the translation. Several conditions were kept constant across all systems to ensure comparability: all systems received the same 120 source sentences, they were set to translate to the same target language variant (European Portuguese), and, in the case of LLM-based translations, the same prompt was used with all models tested.

In the following subsection 3.1, we discuss the selection of the dataset and identify linguistic elements which carry gender information. In subsection 3.2, we describe the annotation process.

3.1 Dataset selection and gender-based structural properties

In this study, we used 120 sentences from WinoMT [18] to test four models with regard to gender bias. WinoMT is a challenge set developed to test gender bias in MT systems, containing a total of 3,888 sentences in English.

The sentences in WinoMT follow a structured pattern, in which they describe a scenario that includes two characters represented by occupation nouns (e.g., *the doctor* and *the nurse*), using a personal pronoun to indicate the gender of one of them (e.g., *he/she/him/her*), for example: *The developer argued with the designer because she did not like the design.*

These sentences are structurally ambiguous with respect to pronoun reference: the pronoun could, in principle, refer to either of the two characters mentioned previously in the sentence. However, for human readers, context and real-world knowledge typically give rise to a preferential interpretation. MT systems and LLMs face a particular challenge in this regard, as they must produce a grammatically marked translation without the pragmatic and cultural resources that guide human comprehension. Although in the sentences of this corpus it is usually only the gender of one of the characters that is indicated by means of the pronoun, due to the asymmetry in the expression of grammatical gender between English and Portuguese (as introduced in section 2.2), when translating from English to Portuguese, the model necessarily has to decide the gender of both characters.

This makes WinoMT a sensitive instrument for detecting gender bias: when a model defaults to a stereotypical gender assignment rather than following the cue provided by the context, the bias becomes directly observable in the translation output.

In this work, 120 sentences were randomly sampled from WinoMT for use as source texts in this study. This subset was chosen to conduct a preliminary evaluation of gender bias in English-to-Portuguese AT, with the aim of identifying relevant patterns and key influencing factors to inform more extensive future evaluations.

Using the occupational stereotype framework from Levy et al. [9], which has been annotated and human-validated, we determined the stereotype directionality of the gender-defined occupation nouns in the selected 120 sentences. In all these sentences, the occupation referred to via gendered pronouns is anti-stereotypical (e.g., *The mechanic gave the clerk a present because she won the lottery*, where *mechanic* is the antecedent of *she*). The 120 sentences consist of 60 pairs of sentences, yielding a balanced distribution of 60 male-targeted and 60 female-targeted instances. The full list of occupations involved is provided in Appendix A.

3.2 Annotation

In the annotation step, we annotated both the source texts and the target texts.

For the source texts, in addition to annotating the gender stereotype of profession nouns occurring in our data, we also annotated nouns whose gender is not defined in the sentence. For example, in the sentence "The mechanic gave the clerk a present because she won the lottery," "the mechanic" is in a co-reference relation with the personal pronoun "she", indicating a female mechanic, which is anti-stereotypical. Meanwhile, the gender of "the clerk" is not determined in the sentence, so we annotated it as an undefined gender noun in this sentence.

Regarding the target texts, we performed fine-grained manual annotation along three dimensions: gender-defined nouns, gender-undefined nouns, and pronouns.

Firstly, for gender-defined nouns, we evaluated whether the grammatical gender in the target text accurately reflected the intended gender of the noun in the source text (e.g., whether "the mechanic" was correctly translated as "*a mecânica*$_{[\text{fem.}]}$"). We annotated the gender in the translation based on the form of the target noun and morphological agreement (e.g., *a professora*/*a gerente* as feminine, *o médico*/*o assistente* as masculine). With the occupational stereotype framework from Levy et al. [9], we also annotated whether the gender of the nouns in the translation preserved the anti-stereotypical nature of the source or reinforced a stereotype.

Secondly, for gender-undefined nouns, we annotated the gender used in the translation and its associated stereotype using the same method.

Finally, we identified gender expression in Portuguese based on grammatical and syntactic cues corresponding to the pronouns in the source texts. We distinguished two primary analytical categories: gender-recoverable and non-gender-recoverable translations. It is worth noting that gender recoverability is not limited to the morphological marking of the occupation nouns themselves, but extends to cases where gender is recoverable from the broader grammatical context.

The gender-recoverable group includes cases where the noun was expressed through a pronoun with gender marking (e.g., *ele*, *ela*, *-o*, *-a*, *este*, *esta*) or via a null subject accompanied by constituents with gender agreement, such as gendered adjectives, allowing us to infer the gender of the noun. In example (2), there are no nouns in the causal clause, but the gender information is made apparent in the subject complement, since *zangada* is a feminine adjective and *stressado* is a masculine adjective and they agree with the gender of the null subject. We also considered gender-marked possessive constructions as gender-recoverable, when these involve contractions of prepositions and personal pronouns (i.e., *dele* and *dela*), see (3).

(2) **(a)** *A advogada gritou com o cabeleireiro porque estava zangada.* (The lawyer$_{[\text{fem.}]}$ yelled at the hairdresser$_{[\text{masc.}]}$ because she was mad.)

(b) *O trabalhador da construção civil marcou uma consulta com o conselheiro porque estava stressado.* (The construction worker$_{[\text{masc.}]}$ scheduled an appointment with the counselor$_{[\text{masc.}]}$ because he was stressed.)

(3) **(a)** *O livro dele está perdido.* (His book is lost.)

(b) *O livro dela está perdido.* (Her book is lost.)

(4) *O cozinheiro preparou um prato para a professora porque tinha fome.* (The cook$_{[\text{masc.}]}$ prepared a dish for the teacher$_{[\text{fem.}]}$ because he was hungry.)

Table 2 Gender-defined nouns: translation accuracy and gender bias.

	Google Translate	DeepL	ChatGPT-4o	DeepSeek-V3
Correct (Anti-stereotype)	81 (67.5%)	70 (58.3%)	66 (55%)	62 (51.7%)
Wrong (Stereotype)	39 (32.5%)	50 (41.7%)	54 (45%)	58 (48.3%)
Gender bias Δ (M% − 50%)	+17.5%	+20.8%	−3.3%	+25%
Total	120 (100%)	120 (100%)	120 (100%)	120 (100%)

Table 3 Mixed-effects logistic regression predicting gender-defined nouns translation accuracy.

Predictor	β	OR	95% CI		p	
Ref.: male target gender, NMT						
Target gender: F	−1.724	0.18	[0.08,	0.38]	<.001	***
System type: LLM	−0.619	0.54	[0.34,	0.86]	.009	**
$N = 480$ AIC = 569.9 $\sigma^2_{\text{sentence}} = 2.33$						

Conversely, the non-gender-recoverable group consists of three types: (i) possessive pronouns that agree with the possessed object rather than the possessor (e.g., *o seu livro*, *a sua casa*, where "*seu*/*sua*" agree with the gender of "*livro*/*casa*", not the referent of a target noun that is the possessor); (ii) personal pronouns in clitic forms lacking morphological gender marking, such as *-lhe* or *-se*, which can therefore have either a male or female antecedent; and (iii) null subjects without any accompanying gender-marked elements, as shown in example (4).

The annotation process was conducted by the first author, proficient in both Portuguese and English, and validated by the second author, a native speaker of Portuguese also proficient in English.

This categorization enables us to quantify and compare how each system preserves, neutralizes, or omits gender information in the target text.

4 Results

In this section, we present and discuss the results obtained in the stages described previously. The analysis is structured into three main parts, section 4.1 examines gender-defined nouns, where the gender of the referent is explicitly indicated in the source sentences. Section 4.2 explores gender-undefined nouns, analyzing cases where gender is not overtly indicated in the source sentence. Section 4.3 focuses on pronouns, investigating their use in translation, and how they contribute to the overall gender representation in discourse.

4.1 Gender-defined nouns

Table 2 shows the results of gender translation when the referent of the target noun is explicitly determined by a gendered pronoun in the source sentence. Google Translate performed best among the systems, with 67.5% of the translations correctly preserving the overtly indicated gender. DeepL followed with 58.3%, while ChatGPT-4o and DeepSeek-V3 trailed at 55% and 51.7% respectively. These results suggest a considerable drop in accuracy from Google Translate to the other systems, highlighting challenges in handling even explicitly marked gender.

Table 4 Gender-undefined nouns translation results.

	Google Translate	DeepL	ChatGPT-4o	DeepSeek-V3
Stereotype	106 (88.3%)	99 (82.5%)	91 (75.8%)	87 (72.5%)
M	60 (50%)	59 (49.2%)	53 (44.2%)	46 (38.3%)
F	46 (38.3%)	40 (33.3%)	38 (31.7%)	41 (34.2%)
Anti-stereotype	14 (11.7%)	21 (17.5%)	29 (24.2%)	33 (27.5%)
M	14 (11.7%)	20 (16.7%)	22 (18.3%)	19 (15.8%)
F	0 (0%)	1 (0.8%)	7 (5.8%)	14 (11.7%)
Δ (M% − 50%)	+11.7%	+15.9%	+12.5%	+4.2%
Total	120 (100%)	120 (100%)	120 (100%)	120 (100%)

Table 2 also presents the directionality of gender translation. Despite all source sentences being anti-stereotypical by design (i.e., feminine pronouns linked to stereotypically male professions and vice versa), all four systems showed a strong preference for translating the target noun as masculine, as reflected in the gender bias Δ. DeepSeek-V3 displayed the most extreme bias ($\Delta = +25\%$), followed by DeepL ($\Delta = +20.8\%$) and Google Translate ($\Delta = +17.5\%$). ChatGPT-4o was the only system approaching gender balance ($\Delta = -3.3\%$), with a slight feminine tendency.

Because all source instances were anti-stereotypical, any correct gender translation results in an anti-stereotypical target output. Therefore, the proportion of stereotype-aligned translations directly reflects the number of gender translation errors. DeepSeek-V3 produced the highest rate of stereotype-conforming outputs (48.3%), followed by ChatGPT-4o (45%), DeepL (41.7%), and Google Translate (32.5%). These results indicate that gender mistranslations are not random, but systematically skew toward reinforcing dominant gender stereotypes, particularly in the form of default masculine rendering of professions.

To further examine the significance of these patterns, a mixed-effects logistic regression was conducted with translation accuracy as the dependent variable, source gender and system type as fixed effects, and sentence as a random effect, shown in Table 3. Results confirmed that target gender significantly predicted translation accuracy (OR = 0.18, 95% CI $[0.08, 0.38]$, $p < .001$), with female-targeted translations being substantially less likely to be correct than male-targeted ones. System type also emerged as a significant predictor (OR = 0.54, 95% CI $[0.34, 0.86]$, $p = .009$), with LLM systems showing lower predicted accuracy than NMT systems overall. The substantial random effect variance ($\sigma^2 = 2.33$) justified the inclusion of sentence as a random effect.

4.2 Gender-undefined nouns

Table 4 presents the translation behavior of the four models when handling gender-undefined nouns, i.e., instances where the source text provides no explicit gender cues, requiring the model to opt for a given gender based only on the noun.

Unsurprisingly, across all systems, stereotype-aligned translations dominate overwhelmingly, ranging from 72.5% (DeepSeek-V3) to 88.3% (Google Translate). All four systems also show a masculine bias, as reflected in the positive Δ values, though DeepSeek-V3 is closest to gender balance ($\Delta = +4.2\%$). Anti-stereotypical translations remain rare across all systems, and feminine anti-stereotypical outputs are almost non-existing in MT: Google

Table 5 Pronoun gender recoverability in the translation.

	Google Translate	DeepL	ChatGPT-4o	DeepSeek-V3
Recoverable	70 (58.3%)	69 (57.5%)	89 (74.2%)	116 (96.7%)
of which M	27 (22.5%)	37 (30.8%)	46 (38.3%)	58 (48.3%)
of which F	43 (35.8%)	32 (26.7%)	43 (35.8%)	58 (48.3%)
Unrecoverable	50 (41.7%)	51 (42.5%)	31 (25.8%)	4 (3.3%)
Total	120 (100%)	120 (100%)	120 (100%)	120 (100%)

Table 6 Pronoun gender translation accuracy by gender.

	Google Translate	DeepL	ChatGPT-4o	DeepSeek-V3
Overall (out of 120)				
Correct	65 (54.2%)	65 (54.2%)	86 (71.7%)	116 (96.7%)
Wrong	5 (4.2%)	4 (3.3%)	3 (2.5%)	0 (0%)
Unrecoverable	50 (41.7%)	51 (42.5%)	31 (25.8%)	4 (3.3%)
Male-defined pronouns (out of 60)				
Correct (M to M)	26 (43.3%)	35 (58.3%)	45 (75%)	58 (96.7%)
Wrong (M to F)	4 (6.7%)	2 (3.3%)	2 (3.3%)	0 (0%)
Unrecoverable	30 (50%)	23 (38.3%)	13 (21.7%)	2 (3.3%)
Female-defined pronouns (out of 60)				
Correct (F to F)	39 (65%)	30 (50%)	41 (68.3%)	58 (96.7%)
Wrong (F to M)	1 (1.7%)	2 (3.3%)	1 (1.7%)	0 (0%)
Unrecoverable	20 (33.3%)	28 (46.7%)	18 (30%)	2 (3.3%)
Total	120 (100%)	120 (100%)	120 (100%)	120 (100%)

Translate produces none, DeepL only 1 case (0.8%). These remain particularly scarce in LLM-produced translations: DeepSeek-V3 produces the highest at 14 cases (11.7%) and GPT-4o generates 7 cases (5.8%).

Overall, these results point to a systemic default toward masculine and, as to be expected, stereotype-conforming translations in the absence of explicit gender markers. While the two LLMs show marginally more balanced outputs compared to the NMT systems, the overwhelming majority of outputs across all systems reinforce stereotyped gender-occupation associations, highlighting persistent bias in undefined gender resolution.

4.3 Pronouns

In this section, we will discuss the translation outputs of the pronoun occurring in the source texts, which is responsible for whether the pronoun is rendered indicating the gender of the target noun. We divided the outputs into two main cases: gender information is recoverable in the translation, in which case we can determine whether the pronoun is masculine or feminine; and gender information is not recoverable in the translation.

Since Portuguese allows for null subjects and various pronominal forms that may or may not encode gender, whether the gender information of the source pronoun is preserved or lost in translation directly reflects whether the model has correctly resolved the referent of the pronoun, making it a meaningful dimension of evaluation alongside the analysis of the occupation noun itself.

Table 5 presents the distribution of gender-recoverable and unrecoverable pronoun translations across outputs produced by the four models. DeepSeek-V3 demonstrates the highest recoverability rate (96.7%), with an equal distribution between male and female (both 48.3%), suggesting balanced gender preservation. ChatGPT-4o follows with 74.2% recoverability and a relatively even gender balance. In contrast, Google Translate and DeepL lag behind with recoverability rates of 58.3% and 57.5% respectively. Interestingly, Google Translate recovers female referents more often (35.8%) than male (22.5%), while DeepL shows a slight male preference.

Table 6 presents translation accuracy in the translation of pronouns and further breaks it down by gender. Overall, DeepSeek-V3 vastly outperforms other models with a 96.7% rate of correct translations and no wrong gender assignments. GPT-4o follows with 71.7%, while Google Translate and DeepL each achieve just over half of accurate pronoun translations (54.2%). Wrong gender assignments are minimal across all systems, indicating that when systems do assign gender, they mostly do so correctly.

Breaking down pronoun translation accuracy by gender, all models perform better on female-defined pronouns than male-defined pronouns, with the exception of DeepSeek-V3, which maintains a consistently high performance across both genders (96.7%). Google Translate shows the largest gender gap, with only 43.3% correct male-translations of pronouns compared to 65% for female ones. Overall, these results reveal varying degrees of gender information loss across systems, with DeepSeek-V3 demonstrating the strongest capability to preserve explicit gender cues in translation.

5 Conclusion

In this study, we used a subset of the WinoMT dataset, a challenge set for grammatical gender translations, to test and compare two machine translation systems (Google Translate and DeepL Translator) and two Large Language Models (ChatGPT and DeepSeek), for the English-European Portuguese language pair. We analyzed gender rendering in different conditions, to assess how faithfully gender information is conveyed and whether stereotypical patterns emerge or intensify.

The first research question focused on whether the models tested correctly convey the gender information from the source to the target. Given the contrasting linguistic properties of English and European Portuguese, gender information in the source is mainly conveyed by the pronoun anaphorically linked to a gender unmarked noun in the sentence. The discrepancies observed indicate that gender rendering is not uniformly handled across different part-of-speech, and while some systems may perform better with explicit lexical cues, they may struggle with pronoun resolution or vice versa.

Secondly, our results show that translation outputs do indeed reproduce and, in some cases, amplify gender stereotypes. In the translations of gender-defined nouns (Table 2), most models lean toward producing male target nouns in Portuguese, particularly DeepSeek-V3 (75% male) and DeepL (70.8% male), whereas GPT-4o was the only system to produce a more balanced distribution. This male-dominant translation is more evident when considered alongside the stereotype classification in Table 2: all four models produced translations reproducing gender stereotypes in at least one-third of the cases, with DeepSeek-V3 reaching the highest proportion (48.3%).

The trend continues when analyzing gender-undefined nouns (Table 4). In this scenario, all systems overwhelmingly defaulted to gender-stereotypical translations, especially masculine-stereotype. Unsurprisingly, this suggests that without explicit gender cues, the four models tend to rely heavily on stereotypical associations between occupations and gender.

The third research question investigates how the different models tested in this study handle undefined gender nouns from the source to the target. Even if LLMs show a more balanced behavior in handling undefined gender, all models demonstrate a preference for outputting male-centric and stereotypical translations, underscoring the challenge of neutrality when gender is unspecified in the source text.

While the present study offers some insights into gender representation in English-to-Portuguese AT, there are several limitations. Firstly, the dataset includes only 120 source sentences, a subset intentionally selected to provide experimental foundations for larger-scale studies focused on the aspects that prove to be most determinant. While this scope is appropriate for a preliminary investigation, it constrains the statistical robustness and generalizability of the findings. Expanding the dataset to include a broader range of professions, sentence structures, and discourse contexts would strengthen future analyses. Secondly, although the annotation of pronoun translations captured two main categories (i.e., recoverable gender in the translation, and unrecoverable gender), a finer-grained classification of translation strategies could provide deeper insight into model behavior, for example, personal pronouns with or without gender marks, null subjects with or without gender-marked part-of-speech, type of possessive constructions, etc. Lastly, the study analyzed marked gender-defined nouns, undefined-gender nouns, and pronoun translations as separate components. However, a more integrated analysis of how these elements correlate could reveal patterns of system consistency or divergence in handling gender related information in translation.

In conclusion, this research demonstrates that in English-Portuguese AT, it is still challenging for the two MT systems (Google Translate and DeepL Translator) and the two LLMs (ChatGPT-4o and DeepSeek-V3) tested to transmit correctly all the gender information from the source to the target text, and that AT outputs lean towards male-centric and stereotypical gender expression, regardless of whether gender is explicitly defined or undefined in the source, especially in the two MT systems.

Future work could design a larger-scale corpus to incorporate more instances of contexts shown to be challenging in this study, and incorporate pronoun referential tracking to capture the interaction among gender cues inside the sentence. Furthermore, as previously noted, even though the English source sentences in WinoMT are structurally ambiguous with respect to pronoun reference, human translators can resolve this ambiguity based on real-world knowledge and pragmatic reasoning to arrive at a contextually appropriate interpretation. Future work could explore whether these limitations can be mitigated through prompt engineering in LLMs, for instance, by providing explicit instructions about pronoun reference resolution, or through few-shot prompting with human-translated examples that demonstrate gender-accurate, context-sensitive translations. Such approaches may offer a practical pathway toward reducing stereotyped default translations in gender-marked target languages in AT.

References

1 Sofia Bonifácio, Helena Moniz, and Luisa Coheur. Diving into gender translation bias for the portuguese language. In *28th European Conference on Artificial Intelligence, ECAI 2025, including 14th Conference on Prestigious Applications of Intelligent Systems, PAIS 2025*, pages 1067–1074. IOS Press BV, 2025.

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

3 Greville G Corbett. *Gender*. Cambridge University Press, 1991.

4 Marta R Costa-Jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. Interpreting gender bias in neural machine translation: Multilingual architecture matters. In *Proceedings of the aaai conference on artificial intelligence*, volume 36, pages 11855–11863, 2022. doi:10.1609/AAAI.V36I11.21442.

5 Eve Fleisig and Christiane Fellbaum. Mitigating gender bias in machine translation through adversarial learning. *arXiv preprint arXiv:2203.10675*, 2022. doi:10.48550/arXiv.2203.10675.

6 Joel Escudé Font and Marta R Costa-Jussa. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, 2019.

7 Charles F Hockett. *A course in modern linguistics*. The Macmillan Company, 1958.

8 Anna Kibort and Greville G Corbett. Grammatical features inventory. *Typology of grammatical features*, 2008.

9 Shahar Levy, Koren Lazar, and Gabriel Stanovsky. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 2470–2480, 2021. doi:10.18653/V1/2021.FINDINGS-EMNLP.211.

10 Maja Popović and Ekaterina Lapshinova-Koltunski. Gender and bias in amazon review translations: by humans, mt systems and chatgpt. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 22–30, 2024.

11 Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381, 2020. doi:10.1007/S00521-019-04144-6.

12 Randolph Quirk and David Crystal. *A comprehensive grammar of the English language*. Pearson Education India, 2010.

13 Eduardo Buzaglo Raposo, Maria Fernanda Nascimento, Maria Antónia Mota, Luísa Segura, and Amália Mendes. *Gramática do Português*. Fundação Calouste Gulbenkian, Lisboa, 2013.

14 Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. The power of prompts: Evaluating and mitigating gender bias in mt with llms. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139, 2024.

15 Danielle Saunders and Bill Byrne. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 7724–7736, 2020. doi:10.18653/V1/2020.ACL-MAIN.690.

16 Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa 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*, pages 18048–18076, 2024. doi:10.18653/V1/2024.EMNLP-MAIN.1002.

17 Peter Siemund. *Pronominal gender in English*. Routledge London, 2008.

18 Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1679–1684, 2019. doi:10.18653/V1/P19-1164.

19 Katie Wales. *Personal pronouns in present-day English*. Cambridge University Press, 1996.

20 Xiaolan Xu and Sara Mendes. xu-xiaolan/gender-bias-en-pt-at. Dataset, swhId: swh:1:dir:f3d848b15460545b49033fd09dfa2af692d74510 (visited on 2026-07-22). URL: https://github.com/xu-xiaolan/gender-bias-en-pt-at, doi:10.4230/artifacts.27322.

21 Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2979–2989, 2017. doi:10.18653/V1/D17-1323.

A Occupation nouns tested in this study and their gender-stereotype, as retrieved from WinoMT

Male stereotype occupations: analyst, carpenter, CEO, chief, construction worker, cook, developer, driver, farmer, guard, janitor, lawyer, manager, mechanic, physician, salesperson.

Female stereotype occupations: accountant, assistant, auditor, baker, cashier, clerk, counselor, hairdresser, housekeeper, librarian, nurse, secretary, tailor, teacher, writer.