GENDEROUS: Machine Translation and Cross-Linguistic Evaluation of a Gender-Ambiguous Dataset

Janiça Hackenbuchner, Eleni Gkovedarou, Joke Daems
LT[3], Ghent University
Groot-Brittanniëlaan, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

Contributing to research on gender beyond the binary, this work introduces GENDEROUS, a dataset of gender-ambiguous sentences containing gender-marked occupations and adjectives, and sentences with the ambiguous or non-binary pronoun `their`. We cross-linguistically evaluate how machine translation (MT) systems and large language models (LLMs) translate these sentences from English into four grammatical gender languages: Greek, German, Spanish and Dutch. We show the systems' continued default to male-gendered translations, with exceptions (particularly for Dutch). Prompting for alternatives, however, shows potential in attaining more diverse and neutral translations across all languages. An LLM-as-a-judge approach was implemented, where benchmarking against gold standards emphasises the continued need for human annotations.

1 Introduction

Recent advancements in machine translation (MT) and large language models (LLMs) have improved translation quality to so-called near-human performance levels (Popel et al., 2020; Yan et al., 2024). Despite these improvements, systems exhibit gender bias by "systematically and unfairly discriminat[ing] against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum, 1996, p.332). Extensive research shows that MT systems and LLMs continue to struggle with bias and fairness (Zhao et al., 2018; Sun et al., 2019; Savoldi et al., 2021; Kotek et al., 2023), contributing to the discrimination of underrepresented social groups.

Studies on gender bias in MT have predominantly focussed on higher resource languages and binary gender, with few recent studies focussing on gender neutrality (Piergentili et al., 2023; Savoldi et al., 2024), less-represented languages (Sewunetie et al., 2024), and non-binarity in translation (Lardelli, 2023; Chen et al., 2024; Piergentili et al., 2024). Additionally, while most gender bias challenge sets focus on unambiguous sentences, we need a better understanding of MT translation for gender-ambiguous sentences to mitigate MT gender bias (Saunders and Olsen, 2023).

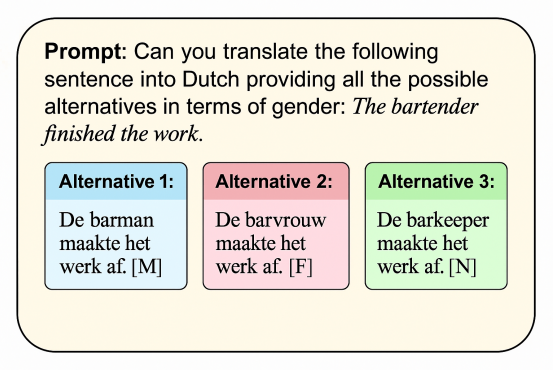

Figure 1: Example illustrating Dutch gender translation alternatives provided by GPT-4o upon being specifically prompted, including annotated gold labels.

To address these research gaps, this paper introduces GENDEROUS[1], a dataset of English gender-ambiguous sentences – constructed without grammatical gender cues, and reports how MT systems and LLMs inherently translate these into four grammatical gender languages: Greek (low resource), German (high resource), Spanish (high resource) and Dutch (medium resource). We analyse how the translations differ in terms of gender for the four target languages and evaluate the extent to which these systems and languages continue to default to male translations [**RQ1**]. We further investigate whether the stereotypicality of an occupational noun, the presence of a gender-inflected adjective, or the interplay between both influence the gender assignment in the translations [**RQ2**] as well as what the impact of the presence of the pronoun `their` is [**RQ3**] – which could be considered

[1] https://github.com/jhacken/GENDEROUS

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 302–319

ambiguous or as a reference to non-binary individuals. We further explore the impact of prompt-based interventions designed to elicit gender-alternative translations from LLMs. This allows us to assess whether prompting increases gender diversity or neutrality in output [**RQ4**]. Given the time-consuming nature of human gender annotations, we additionally explore the potential of LLMs for automatic gender annotation by comparing LLM labels with the human-annotated gold standard [**RQ5**].

2 Related Research

Gender Bias in Occupations & Adjectives Gender bias is a prevalent issue in MT often manifested through occupational stereotypes, i.e. the association of certain occupations with gender (e.g., *nurse* → feminine, *mechanic* → masculine). Occupation-specific bias mirrors real-world employment statistics (Rudinger et al., 2018), with male gender disproportionately linked to the STEM field and high-status roles (Cheryan et al., 2016). MT systems reinforce societal stereotypes by opting for the generic masculine in gender-ambiguous contexts (Schiebinger, 2014; Savoldi et al., 2021), or by translating more accurately for sentences involving men as the training data naturally feature men more than women (Saunders and Byrne, 2020). This phenomenon becomes particularly evident in translations between notional gender languages (e.g., English or Danish), where gender is not always defined, into grammatical gender languages (e.g., German, Italian, or French), where gender needs to be marked in most utterances (Currey et al., 2022).

Prior work has mainly evaluated occupational gender bias in translation using coreference test suites such as WinoMT (Stanovsky et al., 2019) – comprising WinoGender (Zhao et al., 2018) and WinoBias (Rudinger et al., 2018) – in which each sentence contains a primary entity, referred to with an occupational noun, which is co-referent with a pronoun[2]. Troles and Schmid (2021) extended this dataset by combining occupations with gender-stereotypical adjectives[3] and verbs[4]. These word types usually further compound occupation-specific bias, as they also carry gender bias in their word-embeddings (Bolukbasi et al., 2016; Garg et al., 2018; Basta et al., 2019; Troles and Schmid, 2021). Our work focusses exclusively on the interplay between stereotypical occupations and adjectives in translation. Unlike prior studies, we deliberately avoid pronoun co-reference to examine how certain (biased) lexical items shape gender assignments in exclusively gender-ambiguous cases.

Language Comparison Studies on gender bias in MT and LLMs have primarily examined translations from English into high-resource grammatical gender languages like German, French, and Spanish (Isabelle et al., 2017; Currey et al., 2022; Lardelli et al., 2024; Sant et al., 2024; Piazzolla et al., 2024; Vanmassenhove, 2024; Zhao et al., 2024; Lee et al., 2024). Lower-resource languages such as Arabic (Currey et al., 2022), Ukrainian (Stanovsky et al., 2019), or Polish (Kocmi et al., 2020) – among others – have also been included in evaluation studies; however, two languages have received minimal attention: **Dutch**, evolving from a language with three grammatical genders (masculine/feminine/neuter) to a common/neuter gender system with emerging gender-neutral pronouns (Decock et al., 2025), and **Greek**, whose deeply embedded grammatical gender presents particular challenges. Research shows that Dutch word embeddings contain gender bias (Mulsa and Spanakis, 2020) and that bias is present in LLM output for story generation (Butter, 2024). To the best of our knowledge, this is the first study to explore MT gender bias for Dutch. Greek gender bias research saw a preliminary exploration in Karastergiou and Diamantopoulos (2024)'s analysis of document-level outputs, followed by Mastromichalakis et al. (2024)'s labor-domain bias analysis via knowledge graphs, and Gkovedarou et al. (2025)'s comprehensive sentence-level study, which introduced a controlled evaluation of occupational terms and adjective interactions across Google Translate, DeepL, and GPT-4o. Our work provides a systematic comparison across both well-documented languages (German, DE & Spanish, ES) and understudied ones (Dutch, NL & Greek, EL).

LLMs & Prompting Strategies Due to their remarkable performance across a variety of natural language processing (NLP) tasks, LLMs have been tested for their translation capabilities. Research shows that they can perform on par with or better than some state-of-the-art MT models, mainly due to the fact that their outputs can be controlled through explicit zero- or few-shot prompt-

[2] e.g., *The [developer] argued with the [designer] because [she] did not like the design.*

[3] e.g., *The [sassy] [cook] prepared a dish for the [teacher] because [she] just learned a new dish.*

[4] e.g., *The [receptionist] crochets potholders.*

ing (Moslem et al., 2023; Peng et al., 2023; Rarrick et al., 2023; Sánchez et al., 2024; Lee et al., 2024; Koshkin et al., 2024). At the same time, these systems often reinforce gender stereotypes due to inherent biases in their training data (Kotek et al., 2023; Bas, 2024; Zhao et al., 2024). Work on prompting strategies for gender show that LLMs do not reliably produce multiple or correct gender alternatives (Vanmassenhove, 2024) and that LLMs struggle to correctly translate the gender-ambiguous English `their` from and into lower-resource languages (Ghosh and Caliskan, 2023). Our work evaluates both minimal (relying on LLMs' default behaviour) and controlled prompts (directing gender output) to assess the potential for the reduction of bias.

LLM as the Annotator For the evaluation of the translations, we refrain from using automatic evaluation metrics such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006), as they fail to adequately capture certain linguistic phenomena, such as gender bias (Sennrich, 2017). Instead, we rely on manual evaluation to ensure accurate assessment of the outputs. Scaling this can be costly, though; thus, we also explore the *LLM-as-a-judge* paradigm, which has shown promising agreement with human judgments (Kocmi and Federmann, 2023; Kumar et al., 2024) and has recently been successfully implemented to evaluate gender neutral translations (Piergentili et al., 2025). As these systems may inherit and amplify the very biases they are meant to evaluate, including both gender stereotypes and methodological biases like positional preference (favoring the first option in pairwise comparisons) (Wang et al., 2024; Li et al., 2024), we compare LLM annotations with gold-standard human annotations and calculate inter-annotator agreement (IAA) to determine the reliability of LLMs as annotators.

3 Methodology

3.1 Data Collection

With a focus on gender ambiguity, we compiled **GENDEROUS**, a handcrafted dataset[5] of sentences specifically including statistically stereotypical occupational nouns and gender-inflected adjectives. To this end, we selected 30 occupational nouns as listed in Troles and Schmid (2021) taken from US Labor Statistics[6], and ensured that they still coincide with the most recent statistics from 2024. Among these 30 occupations, ten were the top female-dominated occupations, ten were the top male-dominated occupations, and ten occupations were relatively 'neutral', held by both men and women. Starting with these occupations, we compiled 30 **base** sentences.

Gender-inflected adjectives were taken from Charlesworth et al. (2021), who measured the gender-inflection in word embeddings of adjectives. From their list, we chose the top five female-inflected adjectives, the top five male-inflected adjectives and the top five neutral (neither male- nor female-inflected) adjectives. We combined each occupational noun with each gender-inflected adjective, resulting in 15 sentences per noun and a total of 450 sentences.

Additionally, we re-formulated the 30 base sentences to explicitly include the pronoun `their` – simultaneously ambiguous and a reference to non-binary individuals.

Our final handcrafted dataset contains **510** English ambiguous source sentences and their respective parallel translations into Greek, German, Spanish, and Dutch, and consists of (i) 30 base sentences, e.g., *The assistant finished the work*, (ii) 30 `their` sentences, e.g., *The assistant finished their work*, and (iii) 450 sentences including adjectives, e.g., *The clever assistant finished the work.*

3.2 Translation Generation

The dataset was translated into Greek, German, Spanish, and Dutch using two MT systems, DeepL[7] and Google Translate[8], and two LLMs, GPT-4o[9] (`gpt-4o-2024-11-20`) and EuroLLM 9B[10]. The MT translations were done using the DeepL and Google APIs in February 2025. To run translations with the two LLMs, we tested two straightforward prompting strategies, as done in Vanmassenhove (2024), and ran these in March 2025. We ran the first prompt to test how these systems translate gender-ambiguous sentences for the different languages by default and compared these results to the MT translations. Prompt 1: *Can you translate the following sentence into [target language]:*

[5] All data and code is publicly available: https://github.com/jhacken/GENDEROUS

[6] https://www.bls.gov/cps/cpsaat11.htm

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

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

[9] https://chatgpt.com/

[10] https://huggingface.co/utter-project/EuroLLM-9B-Instruct

{input_text}.

To analyse what other translation variations in terms of gender the LLMs could provide, we ran a second prompt on the 30 base sentences for the four languages. This prompt explicitly instructed the LLM to produce additional translation possibilities. Prompt 2: *Can you translate the following sentence into [target language] providing all the possible alternatives in terms of gender: {input_text}*. This analysis is covered in Section 4.2.

EuroLLM exhibited strong limitations in responding to this second prompting strategy. Upon being prompted to provide translations with all possible gender alternatives, EuroLLM kindly responds *"It's great that you're aware of the importance of considering gender when translating sentences"*, but does not end up providing useful translations, if any. Most outputs were missing, cut off halfway, or provided in English. The LLM provided strange results such as "The paralegal finished theirs work" in English (untranslated and misspelled), or provided a Dutch inaccurate explanation about the 'salesperson' being male. Therefore, the results will only discuss the outputs of GPT-4o, which aligns with findings in Piergentili et al. (2025), where GPT-4o performs best on evaluation.

3.3 Evaluation

The evaluation of these translations depends on how the occupational noun is gendered in the target languages. To this aim, human-annotated gold labels for each translation are provided by the authors of this paper, who have a native or C2 competence in their assigned target language. The gold labels were assigned to each translation by manually providing a label of `F`, `M`, `N` or `error` for female, male, neutral or error, respectively. Every translation in each parallel dataset, therefore, comes with a gold label in terms of gender in the translation.

A translation was labelled as `error` if the translation was incorrect (e.g., incorrect translation of an adjective, [EL] "Ο διατροφολόγος του στάβλου τελείωσε τη δουλειά.", or an incorrect translation of the noun, [NL] "De saaie paralegal maakte het werk af.") or if genders were mixed (e.g., incorrect agreement [ES] "El serio ama de llaves terminó el trabajo." or [DE] "Der fleißige Reinigungskraft hat die Arbeit erledigt."). A translation was labelled as `N` (neutral) if the person referred to could be of any gender (e.g., [NL] "De slimme assistent maakte het werk af." or [DE] "Die angenehme Reinigungskraft beendete die Arbeit." or if the translation of a `their` sentence includes a non-binary pronoun (e.g., 'hun' [NL] as in "De chauffeur beëindigde hun werk.").

Moreover, we tested the LLM-as-a-judge approach and whether an LLM is capable of correctly evaluating the gender of the occupational noun in the translations by prompting GPT-4o to assign the above labels to the outputs (Appendix A.1). We benchmark the human gold labels against the LLM annotations and present an inter-annotator agreement between both (Section 4.3).

4 Results

4.1 Gender in Translations

System & Language Comparisons [RQ1] Figure 2 depicts a complete overview of how the systems translated the gender-ambiguous sentences (excluding the 30 `their` sentences) into the four languages in terms of gender. The heatmaps show that the majority of sentences were translated into **male** for Greek, German and Spanish by both the MT systems and LLMs.

Overall, EuroLLM provided the lowest number of male translations, with an average across these three languages of 65.67%. While it would be tempting to interpret this as a lower male bias in the output, this is partially due to errors in the output. In comparison, GPT-4o had the highest number of male translations with an average of 76.67%. The MT systems, DeepL and Google Translate, had a relatively similar number of male translations across these three languages, with an average of 70.33% and 71.33% respectively.

Spanish had the overall highest number of male translations with 80% translated by GPT-4o and an average across the systems of 73.5%. Greek had the second most male translations with an average across the systems of 72.25%, while German had the lowest number of male translations with an average across all systems of 67.25%.

Dutch is an outlier, where the majority of sentences were translated into neutral by all systems. While occupational nouns in Dutch can be grammatically masculine, these are most often used generically, similarly to English. Most Dutch occupational nouns do not have a grammatically feminine equivalent, and those that do sometimes have a different connotation (e.g., the word *boer* (farmer) has a feminine variant *boerin*, but this is most often interpreted as 'a farmer's wife'). For the purpose of

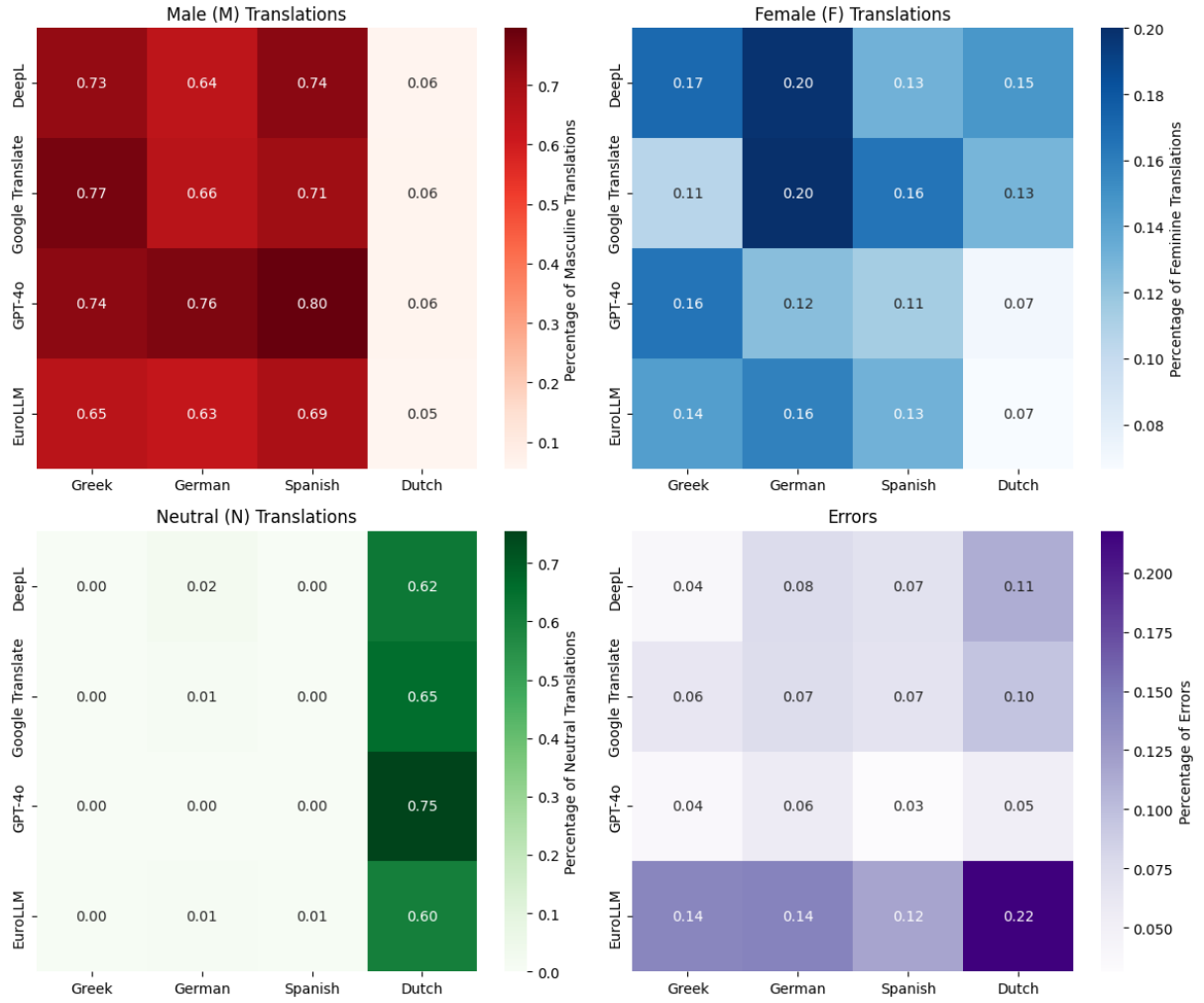

Figure 2: Heatmap comparing the gendered outputs in translation across all languages and systems for Prompt 1 for all sentences, excluding the 30 `their` sentences.

this analysis, nouns that received the 'm/v/x' label in Van Dale[11] (the main dictionary of the Dutch language) were labeled as 'neutral' rather than 'male'. Differences can be seen here where, with 75%, the most sentences were translated as neutral by GPT-4o, and, with 60%, the fewest neutral translations were provided by EuroLLM. As opposed to Dutch, the systems provided no neutral translation for Greek or negligible neutral translations for Spanish and German.

Dutch translations from EuroLLM led to fewer 'neutral' translations because it had the highest error rate of 22%. Overall, EuroLLM had the highest error rates across the four languages. In comparison, GPT-4o had the lowest overall error rates, with a maximum of 6% for German. This value is mostly due to the fact that the occupational noun 'paralegal' was predominantly and erroneously translated as *Paralegal* in German.

In comparison to the LLMs, the two MT systems provided a higher number of female translations across the four languages. Here, German had the highest number of female translations, with 20% as translated by DeepL and Google Translate, and an overall average of 17% across the four systems. Dutch had the lowest number of female translations, with an overall average of 10% across the four systems. Once again, this is due to the combination of higher error rates for EuroLLM and more translations in Dutch being neutral.

Influence of Adjectives [RQ2] As the previous analysis showed Dutch to be an outlier, we limit the adjective analysis to the other three languages. Regardless of adjective type, masculine and neutral occupation terms were practically always translated

[11] https://www.vandale.nl/

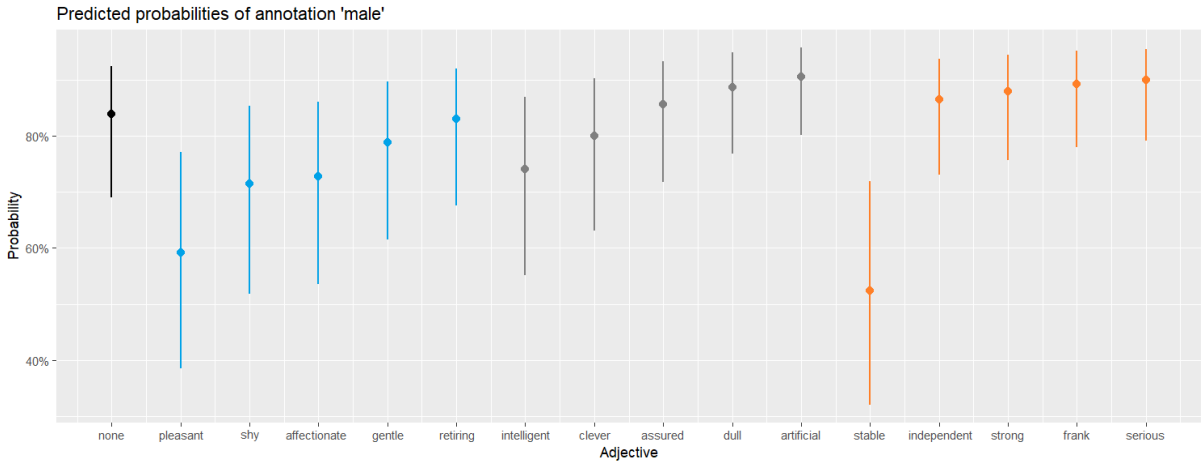

Figure 3: Probabilities of sentences with female-dominated occupation terms containing specific adjectives being annotated as 'male'. Black = base-sentence without adjective, blue = female-inflected adjectives, grey = neutral adjectives, orange = male-inflected adjectives.

as masculine (87%-93%), with up to 9% errors and only rare instances of neutral or female translations. More variation could be observed for the female occupation terms, so the following analysis was conducted on female occupation terms only. We ran a multiple binary logistic regression using the glm function from the stats package in R (R Core Team, 2024) to check if the system, language, occupation term, and adjective would lead to differences in probability of a sentence being annotated as 'male'. All four potential predictors with interaction effects were compared and the best fitting model (based on lowest AIC value) was retained. The final model included all predictors with an interaction effect for language and occupation term. The DHARMa package (Hartig, 2024) was used for residual diagnostics and showed no problems.

Model summary confirmed that GPT-4o increased the probability of a 'male' translation. With regards to specific nouns, 'cleaner', 'dietitian', 'housekeeper', 'nutritionist', 'paralegal', 'receptionist', and 'secretary' were all less likely to be translated as male than 'assistant' (Appendix A.3.2). Among the female-stereotypical occupation nouns, 'assistant' was the occupation least represented by women in the US, which highlights the overlap between an occupation's gender representation (in the US job market) with how this occupation is translated by MT systems.

Compared to sentences without an adjective (Figure 3), sentences with the female-inflected adjectives 'pleasant' and 'shy' were less likely to be annotated as male. The male-inflected adjective 'stable' was the least likely to be translated as male, which, however, was likely due to the high number of erronous translations ('stable' was often translated in relation to horses). Interaction effects showed that 'dietitian' is more likely to be male in Greek compared to German, and 'paralegal' is more likely to be male in Spanish compared to German (again likely error-related), whereas 'receptionist' is less likely to be male in Spanish compared to German. The full model summary and odds ratio with confidence intervals is depicted in Appendix A.3.3 and A.3.4. Overall, translations of sentences with female occupation nouns were marked female most if in combination with a female-inflected adjective and least if in combination with a male-inflected adjective (Appendix A.3.1).

`Their` Sentences [RQ3] While binary gendered translations were deemed acceptable for the translation of the base sentences (Figure 2), these were considered incorrect for the `their` sentences[12] (as the person's gender referred to is either non-binary or unknown). Here, systems should produce a neutral translation. Table 1 shows the gender these 30 sentences were translated into. For example, for Greek, 76% of `their` sentences were translated as male by DeepL. The colours indicate how these percentages differ from those for the base sentences[13], with green values (with border) indicating a desirable change and red values (no border) an

[12] e.g., *The assistant finished **their** work.*
[13] e.g., *The assistant finished **the** work.*

System	EL	DE	ES	NL
Male label				
DeepL	.76	.53	.77	.70
GT	.50	.53	.83	.73
GPT-4o	.87	.83	.87	.23
EuroLLM	.87	.57	.73	.63
Female label				
DeepL	.17	.27	.17	0
GT	.07	.20	.13	.20
GPT-4o	.10	.17	.13	.03
EuroLLM	.07	.27	.20	.17
Neutral label				
DeepL	0	.07	0	.07
GT	0	.03	0	0
GPT-4o	0	0	0	.70
EuroLLM	.03	.07	0	.17
Errors label				
DeepL	.07	.13	.07	.23
GT	.43	.23	.03	.07
GPT-4o	.03	0	0	.03
EuroLLM	.03	.10	.07	.03

Table 1: Label distribution for translations of the `their` sentences across systems and languages. The values are represented as decimal percentages.

undesirable change. For the 'male' and 'female' labels, the desirable change would be a reduction, as binary gendered labels are considered incorrect. For the 'neutral' labels, a desirable change would be an increase.

However, Table 1 shows that this was never the case. The number of 'neutral' labels either remained the same (around 0) for Greek, German, or Spanish, or considerably decreased for Dutch, where `their` was often incorrectly translated as 'his' ('*zijn* werk' instead of the neutral '*het* werk', as in "De kapper is klaar met zijn werk.") or as 'her' ('*haar* werk', as in "De huishoudster was klaar met haar werk").

Instead, the number of errors have noticeably increased in almost every scenario. A frequent source of errors by MT systems for German, and to a lesser extent for Spanish, was that, triggered by the `their` pronoun, the singular person was often mistakenly translated as plural (e.g., where "The firefighter finished their work" was translated as [DE] "Die Feuerwehrleute haben ihre Arbeit beendet" or as [ES] "Los bomberos terminaron su trabajo"). Interestingly, this did not occur for either of the LLMs. Dutch and Greek, on the other hand, had errors where there was an incorrect agreement in the translation (e.g., "Ο βοηθός τελείωσε τη δουλειά τους." or "De diëtiste maakte hun werk af.").

4.2 Prompt for Alternatives [RQ4]

Table 2 provides an overview of the number of gender-alternative translations provided by GPT-4o on the basis of Prompt 2 (Section 3.2) for the 30 **base** sentences. The most translation alternatives were provided for German with an average of 3.8 translations per input sentence (and a max. of 8). The fewest translation alternatives were provided for Greek, with an average of 2.4 translations per input sentence. On average, most of the translation alternatives were grammatically correct, with Greek and Dutch having the highest percentage of 97% and 99%, respectively. German had the highest error rate of 11% due to the incorrect translation output of *Paralegal*.

Prompt 2 Results: Overview

	# of TRs.	% Correct
EL	2.4	.97
DE	**3.8**	.89
ES	2.6	.93
NL	2.7	**.99**

Table 2: Overview of GPT-4o translation results and their accuracy as alternatives provided in response to Prompt 2. All values are averages across the outputs and across languages.

Table 3 shows which gender the translation alternatives were provided in. Due to the nature of the outputs, the gender in the translations were either annotated as `F`, `M`, `N` or `error` as before, or as `M+F` or as `N/I` if one single translation output included both a male and female form, or a neutral inclusive form, respectively. Examples of outputs are shown in Appendix A.2.1.

In comparison to Prompt 1, where only a single translation was given (predominantly as male), we now have an average of gendered translations across the languages. For Greek, there were even slightly more alternatives marked as female with 50%, while 47% of alternatives were male. Unfortunately, Greek translation alternatives continued to remain in the binary, with only 1% being translated as neutral or neutral/inclusive. In Spanish translation alternatives, we now see a sharp increase in female translations and an additional 8% marked

as either neutral or neutral/inclusive. German translations were now balanced between the binary and experienced an increase to 16% for neutral translations. Dutch, similarly as before, has a lower number of male alternatives, and with 54%, the highest number of neutral translation alternatives.

Prompt 2 Results: Gender Analysis

	M	F	M+F	N	N/I
EL	.47	.50	0	.01	.01
DE	.29	.29	.06	.16	.01
ES	.43	.39	.03	.04	.04
NL	.07	.36	.01	.54	0

Table 3: Overview of GPT-4o translation results and their gender-inflection as alternatives provided in response to Prompt 2. All values are decimal averages across the outputs per language.

Results to Prompt 2 show that the LLM is capable of producing more gender-neutral translations across the evaluated languages, with notable neutral or neutral/inclusive improvements observed in German and Spanish, when compared to the default Prompt 1. In addition to increased neutrality, the LLM also yields a near-equal distribution of male and female translations. These findings suggest that, when appropriately prompted, the model can mitigate its tendency to default to masculine forms by generating more balanced gender representations, including a greater proportion of female and neutral (inclusive) translations.

4.3 Annotation Evaluation [RQ5]

Table 4 shows the inter-annotator agreement (Cohen's Kappa (Cohen, 1960)) between GPT-4o generated gender-annotations and gold-standard human annotations across all four systems and languages. Dutch, again, is an outlier with the lowest IAA across all systems with an overall average of κ=0.14. This is due to the fact that GPT labelled most of the Dutch translations as male, instead of as neutral, as was done for the gold label. We see the overall highest IAAs for Spanish, with a 'moderate' average across the four systems of κ=0.69. The highest IAA for a single system was calculated for Greek, with an 'almost perfect' value of κ=0.85.

Interestingly, the highest IAAs are calculated for the MT systems, with the highest overall IAAs for DeepL, and lower IAAs for the LLMs. EuroLLM has the overall lowest IAA with an average of κ=0.35 (again likely error-related). Even though

IAA Cohen's Kappa (κ)

System	EL	DE	ES	NL
DeepL	.85	.74	.76	.14
GT	.70	.72	.80	.16
GPT-4o	.51	.42	.68	.18
EuroLLM	.36	.18	.52	.09

Table 4: Inter-annotator agreement values, in Cohen's Kappa, between GPT-4o as a Judge and the gold label for gender evaluation in the translations. IAA calculated for all systems and all languages.

GPT could also provide an 'error' label for a translation, it seldom did. Human annotators are seen to be more critical and take grammar and meaning of the sentence as a whole into account.

Overall, only one of the IAAs is 'almost perfect', while seven of the IAAs are above the 'moderate' threshold. Nine of the IAAs are below this threshold, emphasising the continued need for gold-standard annotations.

5 Discussion & Conclusion

This study offers novel insights into gender bias in MT systems and LLMs through: (1) the introduction of GENDEROUS, a handcrafted dataset of gender-ambiguous sentences, (2) the comparison of two lesser researched languages (NL, EL) with more widely investigated ones (DE, ES), and (3) the analysis of non-binary linguistic forms.

RQ1: Differences in gender distribution across systems & languages Our results confirm the persistence of masculine defaults (Bas, 2024) across both MT systems and LLMs for gender-ambiguous occupational terms and reveal how deeply embedded societal stereotypes remain in language technologies, even for *artificially-intelligent* models like GPT-4o. Across systems, EuroLLM produced the most errors. Our findings confirm patterns for German and Spanish established in earlier work, with systems being slightly more inclusive for German than for Spanish (Stanovsky et al., 2019). MT systems produced somewhat more female translations than LLMs, except for Greek, where the distribution was similar across systems. Dutch was shown to be an outlier.

RQ2: Impact of stereotypicality of nouns & adjectives Unlike the findings by Troles & Schmid (2021), where gender in MT translation was strongly influenced by adjective stereotypicality

in coreference scenarios, our findings showed that stereotypical male and neutral occupational nouns were predominantly translated as male, regardless of the types of adjectives in the sentence. This indicates that noun stereotypicality appears to be a stronger gender predictor than adjective stereotypicality in ambiguous gender sentences, supporting the need for more research into MT bias in gender-ambiguous scenarios (Saunders and Olsen, 2023). Merely stereotpyical female nouns were more likely to be translated as female ('assistant' and 'hairdresser' to a lesser extent), particularly in combination with female-inflected adjectives.

RQ3: `Their` sentences The presence of the pronoun `their` has been shown to lead to mistranslations in earlier work on understudied languages (Ghosh and Caliskan, 2023). The `their` sentences in our study were also predominantly translated incorrectly (either as binary gendered or as error) across all systems and languages. This highlights the persistent binary biases exhibited by current technologies, confirming the findings by Lardelli (2023, p.61) that "current MT systems do not recognise non-binary pronouns and erase non-binary idenitities in their outputs". It particularly stresses the need for work on lesser-researched languages such as Dutch, where the presence of the pronoun led to incorrect binary translations, despite the base sentences being translated more neutral.

RQ4: Prompt-based intervention Introducing a tailored prompt (explicitly requesting gender alternatives) led to unusable output for EuroLLM, suggesting that smaller models might struggle with more complex prompts, potentially due to reduced generalisation abilities (Moradi et al., 2024). In contrast with findings by Vanmassenhove (2024), where explicit prompting led to worse results, we noticed that GPT-4o outputs led to better results for the languages studied here. We see that explicit prompting leads to more diversity in output, especially for German, and an occasional (non-binary) gender-inclusive translation for Spanish. It must be noted, however, that the output did not contain systematic strategies, with very different alternatives and suggestions across sentences. Some alternatives were related to the non-gendered elements of the sentence ("finished the work").

RQ5: LLM for gender evaluation After testing GPT-as-a-judge to evaluate gender in translations, overall unsatisfactory inter-annotator agreements with the human gold label show that human annotations continue to be necessary and valuable for in-depth work on gender bias in language technology. LLM annotations for MT output were better than those for LLM output, but results varied widely across languages and systems. The most consistent results were obtained for Spanish, the worst for Dutch. Piergentili et al. (2025), in contrast, find a higher accuracy for LLM evaluation by applying different prompting strategies both on the phrase- and sentence level. However, they equally find evaluation performance to differ across languages.

6 Bias Statement

In this paper, we evaluate English-to-Greek/German/Spanish/Dutch MT and LLM translations. We analyse these systems' default behaviour in translating professional occupations and adjectives, and specifically address the issue of representational harm (Blodgett et al., 2020), categorised into two types: under-representation, which reduces the visibility of certain social groups (such as women and non-binary individuals), and stereotyping, which reinforces negative generalisations (e.g., associating women with less prestigious professions compared to men) (Savoldi et al., 2021).

7 Limitations

Several limitations should be acknowledged. First, the compiled dataset is considered a relatively small size in today's field of research in NLP. This can lead to some very specific issues skewing the results, such as the adjective 'stable' being mistranslated as a horse stable, and the noun 'paralegal' being mistranslated frequently in German and Dutch, leading to errors unrelated to gender specifically. Second, gender bias – particularly in relation to occupational nouns and adjectives – has already been extensively examined in prior work, although usually in combination with coreference resolution. Research into gender-ambiguous sentences is more rare, and ours is the first study to contrast these specific languages. Third, the analysis primarily focusses on default translation behaviour; more advanced prompting strategies were not systematically explored. Especially given the unexpected results for Prompt 2 for EuroLLM, a suggestion for future work would be to more systematically test a variety of prompting strategies in larger and smaller (open-source) models. Equally, more de-

tailed evaluation prompts with a focus on error annotation (as humans tend to be more critical) should be explored in future work. Fourth, Prompt 2 was tested on a limited subset of 30 sentences and on a single system (as EuroLLM results could not be considered for analysis). Finally, previous research similarly focusses on translations from English into grammatical gender languages, whereas a different (notional or genderless) source language and language direction could be of interest. In future work, additional sentences, sentence types, language directions, and prompting strategies could be contrasted across these and additional languages.

Acknowledgments

This study has been partially funded by The Research Foundation – Flanders (FWO), research project 1SH5V24N (from 01.11.2023 until 31.10.2027), and hosted within the Language and Translation Technology Team (LT3) at Ghent University. The computational resources (Stevin Supercomputer Infrastructure) and services partially used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI. Additionally, we would like to thank the anonymous reviewers for their valuable feedback.

References

Hirotugu Akaike. 2011. Akaike's information criterion. *International encyclopedia of statistical science*, pages 25–25.

Tetiana Bas. 2024. Assessing gender bias in llms: Comparing llm outputs with human perceptions and official statistics. *Preprint*, arXiv:2411.13738.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain.

Shiyi Butter. 2024. Unveiling gender bias in occupations: A comparative analysis of gpt-3.5 and llama 2 in the generation of dutch short stories. Master's thesis, Utrecht University.

Tessa E. S. Charlesworth, Victor Yang, Thomas C. Mann, Benedek Kurdi, and Mahzarin R. Banaji. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2):218–240.

Yijie Chen, Yijin Liu, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2024. Beyond binary gender: Evaluating gender-inclusive machine translation with ambiguous attitude words. *Preprint*, arXiv:2407.16266.

Sapna Cheryan, Sianna Ziegler, Amanda Montoya, and Lily Jiang. 2016. Why are some stem fields more gender balanced than others? *Psychological Bulletin*, 143.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sofie Decock, Sarah Van Hoof, Lou-Lou Vanrobaeys, Hanne Verhaegen, and Chloé Vincent. 2025. The dutch gender-neutral pronoun die: more accepted for generic than for specific reference. *Taal & Tongval*, 77(1):76–107.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. 14(3):330–347.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16).

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

Eleni Gkovedarou, Joke Daems, and Luna De Bruyne. 2025. Gender bias in english-to-greek machine translation. *Preprint*, arXiv:2506.09558.

Florian Hartig. 2024. *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.4.7.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Anestis Polychronis Karastergiou and Konstantinos Diamantopoulos. 2024. Gender Issues in Machine Translation. *Transcultural Journal of Humanities & Social Sciences*, 5(1):48–64.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. In *Proceedings of the 5th Conference on Machine Translation (WMT)*, page 357–364. Association for Computational Linguistics.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024. LLMs Are Zero-Shot Context-Aware Simultaneous Translators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1207, Miami, Florida, USA. Association for Computational Linguistics.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24. ACM.

Shachi H Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Manuvinakurike, Nicole Beckage, Hsuan Su, Hung yi Lee, and Lama Nachman. 2024. Decoding biases: Automated methods and llm judges for gender bias detection in language models. *Preprint*, arXiv:2408.03907.

Manuel Lardelli. 2023. Post-editing machine translation beyond the binary: Insights into gender bias and screen activity. In *Translating and the Computer 45*, page 50–64, European Convention Center, Luxembourg. Tradulex.

Manuel Lardelli, Giuseppe Attanasio, and Anne Lauscher. 2024. Building bridges: A dataset for evaluating gender-fair machine translation into German. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7542–7550, Bangkok, Thailand. Association for Computational Linguistics.

Minwoo Lee, Hyukhun Koh, Minsung Kim, and Kyomin Jung. 2024. Fine-grained Gender Control in Machine Translation with Large Language Models. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5416–5430. Association for Computational Linguistics.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024. Split and merge: Aligning position biases in LLM-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108, Miami, Florida, USA. Association for Computational Linguistics.

Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Eva Tsouparopoulou, Dimitris Parsanoglou, Maria Symeonaki, and Giorgos Stamou. 2024. Gost-mt: A knowledge graph for occupation-related gender biases in machine translation. *Preprint*, arXiv:2409.10989.

Milad Moradi, Ke Yan, David Colwell, Matthias Samwald, and Rhona Asgari. 2024. Exploring the landscape of large language models: Foundations, techniques, and challenges. *arXiv preprint arXiv:2404.11973*.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. Proceedings of the 24th Annual Conference of the European Association for Machine Translation, pages 227–237. European Association for Machine Translation.

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in dutch word embeddings. In *Gender Bias in Natural Language Processing*, pages 56–71. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation. Findings of the Association for Computational Linguistics: EMNLP 2023, pages 5622–5633. Association for Computational Linguistics.

Silvia Alma Piazzolla, Beatrice Savoldi, and Luisa Bentivogli. 2024. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems. *Preprint*, arXiv:2306.05882.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83, Tampere, Finland. European Association for Machine Translation.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 300–314, Sheffield, UK. European Association for Machine Translation (EAMT).

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025. An llm-as-a-judge approach for scalable gender-neutral translation evaluation. *arXiv e-prints*.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11:1–15.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. GATE: A Challenge Set for Gender-Ambiguous Translation Examples. pages 845–854. Association for Computing Machinery.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 (Short Papers), pages 8–14. Association for Computational Linguistics.

Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Gender-specific machine translation with large language models. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 148–158, Miami, Florida, USA. Association for Computational Linguistics.

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

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

Danielle Saunders and Katrina Olsen. 2023. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. In *1st Workshop on Gender-Inclusive Translation Technologies*, page 85.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. In *Transactions of the Association for Computational Linguistics*, volume 9, page 845–874, Cambridge, MA. MIT Press.

Beatrice Savoldi, Aandrea Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2024. A prompt response to the demand for automatic gender-neutral translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2: Short Papers, page 256–267. Association for Computational Linguistics.

Londa Schiebinger. 2014. Scientific research must take gender into account. *Nature*, 507(7490):9–9. Publisher: Nature Publishing Group.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Walelign Sewunetie, Atnafu Tonja, Tadesse Belay, Hellina Hailu Nigatu, Gashaw Gebremeskel, Zewdie Mossie, Hussien Seid, and Seid Yimam. 2024. Gender bias evaluation in machine translation for Amharic, Tigrigna, and afaan oromoo. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 1–11, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

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

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 1679–1684.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Jonas-Dario Troles and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine

translation. impact of stereotypical verbs and adjectives. In *Proceedings of the Sixth Conference on Machine Translation*, page 531–541.

Eva Vanmassenhove. 2024. Gender bias in machine translation and the era of large language models. In *Gendered Technology in Translation and Interpreting*, pages 225–252. Routledge.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *CoRR*, abs/2407.03658.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 (Short Papers), pages 15–20. Association for Computational Linguistics.

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

A Appendix

A.1 LLM-as-a-Judge

To benchmark human gold standard annotations, we adopt a straightforward LLM-as-a-judge approach to evaluate gender in the output translations for each language. We provide GPT-4o with the translated text and prompt it for gender labels, for which we apply the following evaluation prompt:

```
The text {input_text} consists of {#} sentences, in which each sentence refers to
one person (e.g., secretary or electrician) in {target_language}. Provide a table
with 2 columns and {#} rows: For each sentence, the first column lists the person
referred to and the second column labels the gender of the person as either female,
male, neutral, or as an error if you see no person or gender (labels: F, M, N, error).
```

A.2 Prompt 2: Example

Source text	Translation Alternatives	Gender
EN	**EL**	
The electrician finished the work.	Ο ξυλουργός ολοκλήρωσε τη δουλειά.	M
	Η ξυλουργός ολοκλήρωσε τη δουλειά.	F
	DE	
	Der Elektriker hat die Arbeit beendet.	M
	Die Elektrikerin hat die Arbeit beendet.	F
	Die Elektrofachkraft hat die Arbeit beendet.	N
	ES	
	El electricista terminó el trabajo.	M
	La electricista terminó el trabajo.	F
	La persona electricista terminó el trabajo.	N
	El o la electricista terminó el trabajo.	M+F
	NL	
	De elektricien heeft het werk afgemaakt.	N
	De elektricien heeft het werk voltooid.	N

Table 5: Example of gender translation alternatives as provided as outputs to Prompt 2 for Greek, German, Spanish and Dutch.

A.2.1 Prompt 2: Gender Alternative Outputs

The LLM partially got very creative. If a translation was clearly erroneous (grammatically or by providing a false translation), it was annotated as an 'error'. However, if the LLM provided a creative, inclusive alternative, we marked it as 'N/I'.

Label	Language	Translation Alternatives
M+F	**DE**	*Der Installateur/die Installateurin hat die Arbeit beendet.*
	ES	*El/la pintor/a terminó el trabajo.*
N	**DE**	*Die Aufsichtsperson hat die Arbeit beendet.*
	ES	*La persona encargada de despachar terminó el trabajo.*
	NL	*De leidinggevende heeft het werk afgemaakt.*
N/I	**EL**	Το καλλιτεχνικό άτομο ολοκλήρωσε το έργο.
	DE	*Der*die Mechaniker*in hat die Arbeit beendet.*
	ES	*Le supervise terminó el trabajo.*
Error	**EL**	Το πρόσωπο που καθαρίζει τελείωσε τη δουλειά.
	DE	*Der Paralegal/die Paralegalin hat die Arbeit abgeschlossen.*
	ES	*La bartender terminó el trabajo.*

Table 6: Example of annotated labels (M+F, N, N/I, error) of gender translation alternatives as provided as outputs to Prompt 2.

A.3 Influence of Gender-Inflected Adjectives

A.3.1 Translations of Female Occupation Nouns

Table 7 depicts the distribution of female-marked translations of stereotypical female nouns in combination with either female-, neutral- or male-inflected adjectives. This table shows that across languages (and systems), stereotypical female occupation nouns were mostly translated as female when in combination with a female-inflected adjective. This number decreases slightly when in combination with a neutral-inflected adjective, and decreases most when in combination with a male-inflected adjective. This shows that the translations of gender-ambiguous sentences with stereotypical female occupation nouns are influenced by the interplay of the noun and a gender-inflected adjective. In comparison, stereotypical male and neutral occupation nouns were predominantly translated as male, with no notable influence from the gender-inflected adjective, and therefore not further analysed here.

Female-marked translations w.r.t. adjectives

Language	Female Adj.	Neutral Adj.	Male Adj.
EL	**0.56**	0.47	0.35
DE	**0.61**	0.51	0.40
ES	**0.54**	0.42	0.32
NL	**0.38**	0.33	0.27

Table 7: Female-label distribution for the translation of sentences with stereotypical female nouns in combination with either female-, neutral- or male-inflected adjectives.

A.3.2 Probabilities of female nouns being translated as male

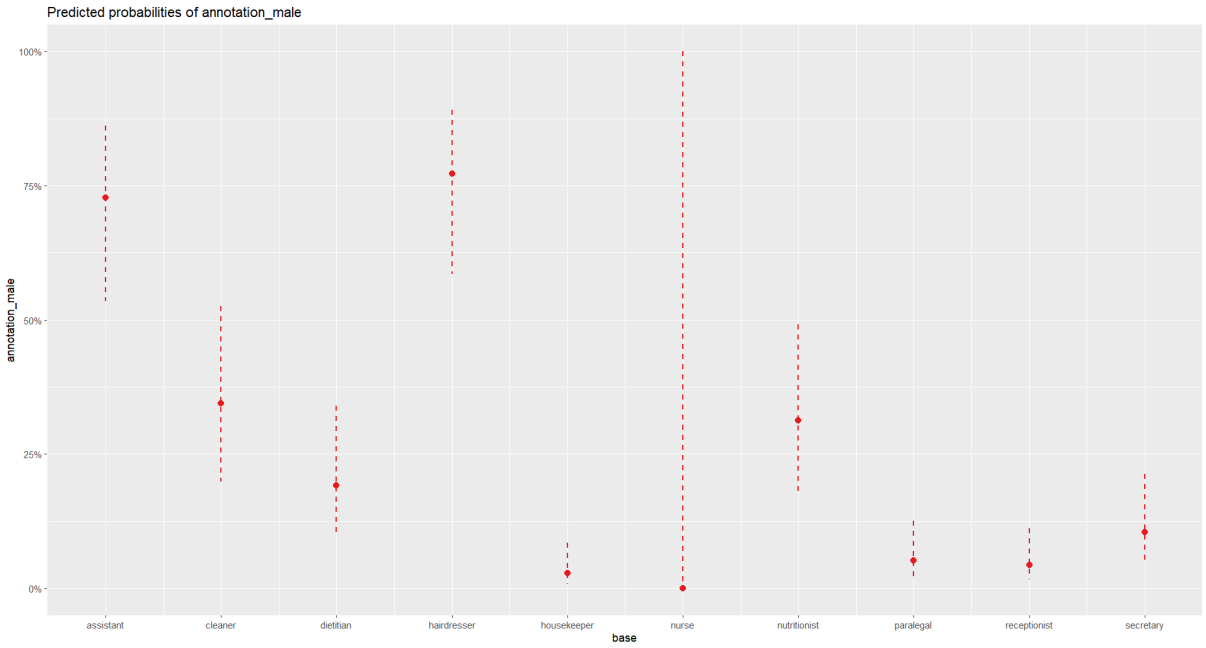

Figure 4: Probabilities of sentences with specific female-dominated occupation terms being annotated as ‘male’. For reference, ‘assistant’ was the noun with the lowest female representation in real world data - 85% of assistants are women, but this explanation does not hold for ‘hairdresser’ with 92%.

A.3.3 Model summary

Predictor	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.65065	0.43392	3.804	0.000142	***
System = EuroLLM	-0.1011	0.18362	-0.551	0.581898	
System = Google Translate	-0.06739	0.18358	-0.367	0.713549	
System = GPT-4o	0.74568	0.18623	4.004	6.23E-05	***
Language = Greek	0.37183	0.50061	0.743	0.457633	
Language = Spanish	0.85981	0.55304	1.555	0.120017	
noun = cleaner	-1.62866	0.42728	-3.812	0.000138	***
noun = dietitian	-2.42399	0.43773	-5.538	3.07E-08	***
noun = hairdresser	0.23863	0.48956	0.487	0.625955	
noun = housekeeper	-4.53427	0.62452	-7.26	3.86E-13	***
noun = nurse	-19.3394	474.711	-0.041	0.967504	
noun = nutritionist	-1.76846	0.42765	-4.135	3.54E-05	***
noun = paralegal	-3.90209	0.53208	-7.334	2.24E-13	***
noun = receptionist	-4.08069	0.55392	-7.367	1.75E-13	***
noun = secretary	-3.12986	0.46713	-6.7	2.08E-11	***
adj = affectionate	-0.66742	0.36659	-1.821	0.068668	.
adj = artificial	0.60936	0.36909	1.651	0.098739	.
adj = assured	0.1341	0.36625	0.366	0.714254	
adj = clever	-0.2671	0.36564	-0.73	0.465087	
adj = dull	0.40426	0.36756	1.1	0.271396	
adj = frank	0.47235	0.36801	1.284	0.199314	
adj = gentle	-0.33376	0.36569	-0.913	0.361405	
adj = independent	0.20136	0.36651	0.549	0.582721	
adj = intelligent	-0.60057	0.36632	-1.639	0.101117	
adj = pleasant	-1.27738	0.37155	-3.438	0.000586	***
adj = retiring	-0.06688	0.36575	-0.183	0.854905	
adj = serious	0.54071	0.36852	1.467	0.142314	
adj = shy	-0.73438	0.36691	-2.001	0.045339	*
adj = stable	-1.55795	0.37567	-4.147	3.37E-05	***
adj = strong	0.33642	0.36716	0.916	0.359521	
Language = Greek:noun = cleaner	-0.44169	0.6248	-0.707	0.479607	
Language = Spanish:noun = cleaner	0.45943	0.68671	0.669	0.503472	
Language = Greek:noun = dietitian	1.38703	0.64217	2.16	0.030779	*
Language = Spanish:noun = dietitian	-0.34489	0.67402	-0.512	0.608869	
Language = Greek:noun = hairdresser	-0.23863	0.7193	-0.332	0.740081	
Language = Spanish:noun = hairdresser	-0.23863	0.79219	-0.301	0.763246	
Language = Greek:noun = housekeeper	-1.82848	1.24402	-1.47	0.141612	
Language = Spanish:noun = housekeeper	-15.6649	474.7115	-0.033	0.973676	
Language = Greek:noun = nurse	12.9766	474.7122	0.027	0.978192	
Language = Spanish:noun = nurse	13.20499	474.7118	0.028	0.977808	
Language = Greek:noun = nutritionist	0.99335	0.64265	1.546	0.122175	
Language = Spanish:noun = nutritionist	-0.07318	0.67199	-0.109	0.91328	
Language = Greek:noun = paralegal	0.29397	0.72636	0.405	0.68568	
Language = Spanish:noun = paralegal	3.41411	0.78438	4.353	1.35E-05	***
Language = Greek:noun = receptionist	0.95674	0.72718	1.316	0.18828	
Language = Spanish:noun = receptionist	-2.77004	1.23294	-2.247	0.024659	*
Language = Greek:noun = secretary	0.00591	0.66418	0.009	0.992901	
Language = Spanish:noun = secretary	-0.85981	0.71612	-1.201	0.229884	

Table 8: Model summary of multiple binary logistic regression. Dependent variable = stereotypically female noun annotated as 'male'. Final model was selected on the basis of lowest AIC value. AIC = Akaike's information criterion (Akaike, 2011), with lower values indicating a lower prediction error, meaning that a model better fits the data. AIC for this particular model was 1570.8, and 2614.1 for the null model without predictors. All possible predictor combinations and interaction effects were tested in order to find the model with the lowest AIC value.

A.3.4 Odds ratio and confidence intervals

predictor	oddsratio	ci_low (2.5)	ci_high (97.5)
System = EuroLLM	0.904	0.630	1.30
System = Google Translate	0.935	0.652	1.34
System = GPT-4o	2.108	1.466	3.04
Language = Greek	1.45	0.546	3.96
Language = Spanish	2.363	0.823	7.43
noun = cleaner	0.196	0.083	0.44
noun = dietitian	0.089	0.036	0.20
noun = hairdresser	1.27	0.486	3.37
noun = housekeeper	0.011	0.003	0.03
noun = nurse	0	0.000	0.00
noun = nutritionist	0.171	0.072	0.39
noun = paralegal	0.02	0.007	0.05
noun = receptionist	0.017	0.005	0.05
noun = secretary	0.044	0.017	0.11
adj = affectionate	0.513	0.249	1.05
adj = artificial	1.839	0.894	3.80
adj = assured	1.144	0.558	2.35
adj = clever	0.766	0.373	1.57
adj = dull	1.498	0.730	3.09
adj = frank	1.604	0.781	3.31
adj = gentle	0.716	0.349	1.47
adj = independent	1.223	0.596	2.51
adj = intelligent	0.548	0.267	1.12
adj = pleasant	0.279	0.134	0.58
adj = retiring	0.935	0.456	1.92
adj = serious	1.717	0.835	3.55
adj = shy	0.48	0.233	0.98
adj = stable	0.211	0.100	0.44
adj = strong	1.4	0.682	2.88
Language = Greek:noun = cleaner	0.643	0.186	2.18
Language = Spanish:noun = cleaner	1.583	0.398	6.00
Language = Greek:noun = dietitian	4.003	1.129	14.15
Language = Spanish:noun = dietitian	0.708	0.182	2.61
Language = Greek:noun = hairdresser	0.788	0.190	3.24
Language = Spanish:noun = hairdresser	0.788	0.164	3.76
Language = Greek:noun = housekeeper	0.161	0.007	1.46
Language = Spanish:noun = housekeeper	0	0.000	0.00
Language = Greek:noun = nurse	432180.6	>1000000	>1000000
Language = Spanish:noun = nurse	543069.2	>1000000	>1000000
Language = Greek:noun = nutritionist	2.7	0.760	9.55
Language = Spanish:noun = nutritionist	0.929	0.239	3.41
Language = Greek:noun = paralegal	1.342	0.323	5.68
Language = Spanish:noun = paralegal	30.39	6.529	144.93
Language = Greek:noun = receptionist	2.603	0.634	11.21
Language = Spanish:noun = receptionist	0.063	0.003	0.53
Language = Greek:noun = secretary	1.006	0.271	3.70
Language = Spanish:noun = secretary	0.423	0.100	1.69

Table 9: Odds ratio with confidence interval for potential predictors of stereotypically female nouns being annotated as ‘male’.