

Gender Bias in English-to-Greek Machine Translation

Eleni Gkovedarou
Ghent University

Eleni.Gkovedarou@ugent.be

Joke Daems
Ghent University

Joke.Daems@ugent.be

Luna De Bruyne
University of Antwerp

Luna.DeBruyne@uantwerpen.be

Abstract

As the demand for inclusive language increases, concern has grown over the susceptibility of machine translation (MT) systems to reinforce gender stereotypes. This study investigates gender bias in two commercial MT systems, Google Translate and DeepL, focusing on the understudied English-to-Greek language pair. We address three aspects of gender bias: *i*) male bias, *ii*) occupational stereotyping, and *iii*) errors in anti-stereotypical translations. Additionally, we explore the potential of prompted GPT-4o as a bias mitigation tool that provides both gender-explicit and gender-neutral alternatives when necessary. To achieve this, we introduce GendEL, a manually crafted bilingual dataset of 240 gender-ambiguous and unambiguous sentences that feature stereotypical occupational nouns and adjectives. We find persistent gender bias in translations by both MT systems; while they perform well in cases where gender is explicitly defined, with DeepL outperforming both Google Translate and GPT-4o in feminine gender-unambiguous sentences, they are far from producing gender-inclusive or neutral translations when the gender is unspecified. GPT-4o shows promise, generating appropriate gendered and neutral alternatives for most ambiguous cases, though residual biases remain evident. As one of the first comprehensive studies on gender bias in English-to-Greek MT, we provide both our data and code at https://github.com/elenigkove/genderbias_EN-EL_MT.

1 Introduction

As the demand for inclusive language grows, the role of technology in shaping linguistic norms becomes increasingly important. While MT is widely used for communication, cost reduction, and accessibility (Nurminen and Koponen, 2020; Vieira,

2020; Asscher and Glikson, 2023) and despite advancements in the field with state-of-the-art neural machine translation (NMT) systems, they often produce inaccurate, ungrammatical, or biased translations, particularly in assigning gender (Stanovsky et al., 2019; Currey et al., 2022). Concern has therefore grown over the susceptibility of those systems to translate based on gender stereotypes and the perpetuation of such biases via them, as they can have real-world, harmful consequences for users and society (Monti, 2020; Savoldi et al., 2021; Lardelli and Gromann, 2023).

A MT model is considered biased “when it *systematically* and *unfairly* discriminates against certain individuals or groups in favour of others” (Friedman and Nissenbaum, 1996). While human translators rely on the wider context to determine the appropriate gender, most current MT systems do not; instead, they rely on spurious correlations in the (biased) training data which often lead to defaulting to either male or female forms (Vanmassenhove et al., 2018; Kocmi et al., 2020). These biases reflect the gender stereotypes that are present in our society. As Saunders et al. (2020) highlight, translations are more accurate for sentences involving men due to the training data naturally featuring men more than women, as well as for sentences that align with stereotypical gender roles. For example, references of “male doctors” are more reliably translated than those of “male nurses” (Sun et al., 2019), while more errors are detected when the source texts exhibit anti-stereotypical professions, e.g. “female doctors” or “male nurses” (Kocmi et al., 2020).

This study focuses on gender bias that occurs when translating from English, a notional gender language where gender is not always specified, into Modern Greek (henceforth Greek), a grammatical gender language where it is morphologically and semantically necessary to mark the gender (Savoldi et al., 2021; Currey et al., 2022). While prior work

© 2025 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

has focused on high-resource languages such as German, Spanish, and French (Currey et al., 2022; Zhao et al., 2024; Lee et al., 2024; Lardelli et al., 2024), Greek remains understudied despite preliminary exploration on document-level (Karastergiou and Diamantopoulos, 2024) and labor-domain bias analysis (Mastromichalakis et al., 2024).

These cross-linguistic differences can lead to ambiguities that are difficult to resolve, especially for sentence-level MT systems (Vanmassenhove et al., 2021), making it more complex to accurately assign gender or maintain gender neutrality. This raises important questions as to what constitutes appropriate handling of gender-ambiguous inputs. Although individual translations (e.g., translating “the worker” as either feminine or masculine) may be grammatically valid, aggregate patterns reveal systemic biases. We argue that ideal MT systems should recognise when gender cannot be determined from context and provide either gender-neutral options, or a range of plausible gendered interpretations. This framework informs our evaluation of existing systems and our proposed LLM-based mitigation approach.

The main contribution of this work includes:

1. We release **GendEL**, the first (handcrafted) dataset for evaluating English-to-Greek translations, which comprises *i*) 240 gender-ambiguous and unambiguous English sentences, *ii*) human-generated Greek alternate translations (feminine/masculine/neutral variants), and *iii*) non-binary representations (singular ‘they’).
2. We focus on the under-represented English-to-Greek language pair and demonstrate that gender bias is persistent in translations by **Google Translate**¹ and **DeepL**.²
3. We explore the potential of prompting a large language model (LLM), **GPT-4o**³ (OpenAI, 2024), to help in bias mitigation by generating not only accurate gender assignments, but also gender-explicit and gender-neutral alternatives for ambiguous English sentences.

Bias Statement In this paper, we analyse English-to-Greek MT outputs to study stereotypical gender associations with professional occupations and adjectives. We analyse the presence of

three patterns of gender bias and specifically address the issue of **representational harm** (Blodgett et al., 2020). Representational harm is 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).

2 Related Work

2.1 Greek as a grammatical gender language

Understanding Greek’s grammatical gender system is important for analysing gender bias in translation, as the language’s structural requirements and gender-inclusive techniques influence the manifestation of gender in the outputs. Greek has three declensions: masculine, feminine, and neuter. Masculine and feminine typically mark human referents, while neuter is mostly assigned to inanimate objects, though certain neuter nouns such as “το άτομο” (the individual) or “το μέλος” (the member) refer to humans. In general, nouns denoting male human beings are grammatically masculine and nouns denoting female human beings are grammatically feminine (Pavlidou et al., 2004). Alvanoudi (2015) points out that Greek’s deeply embedded grammatical gender is restrictive, as gender marking must be encoded in most of the speakers’ utterances.

In this study, the focus is primarily on occupational nouns, which often have overt gender marking through suffixes, e.g. “δάσκαλος” (teacher [M]) and “δασκάλα” (teacher [F]). However, gender marking can also be covert, known as *common gender* or *epicene* nouns, which share the same form for male and female referents. In these cases, disambiguation relies on articles or other modifiers; for instance, “ο/η δικηγόρος” (the [M/F] lawyer) uses a clearly masculine suffix for either gender, while the gender is clarified only by the article. The morphological formation and choice of suffix for occupational nouns can be semantically linked to implicit connotations and is an indication of linguistic sexism (Σαρρή-Χασάν, 2024). Several proposals (Τριανταφυλλίδης, 1963; Τσοπανάκης, 1982; Τσοχαλίδου, 1996; Γκασούχα and Γεωργαλίδου, 2018; Σαρρή-Χασάν, 2024) have been made to feminise such epicenes in ways that align with the morphological and inflectional system of Greek while adhering to grammatical gender agree-

¹<https://translate.google.com/>

²<https://www.deepl.com/en/translator/>

³<https://openai.com/index/hello-gpt-4o/>

ment, e.g. “η δικηγόρινα”. However, despite a slow increase in acceptance and usage, most of these feminised terms have not become standardised in official language use.

In some instances, both masculine and feminine forms exist for occupational nouns, but the feminine versions often carry semantic, stylistic, or register differences. These differences can potentially result in negative connotations or reduced social weight. For example, “δήμαρχος” (mayor) and “δήμαρχέσσα” (the wife of a mayor or a female mayor) differ in both gender marking and societal implications (Kalfadopoulou and Tsigou, 2022).

2.2 Gender-inclusive practices in Greek

Pavlidou et al. (2004) found that Greek vocabulary is male-dominated, with masculine nouns for human reference nearly doubling feminine ones. This male bias is evident also in the use of the generic masculine, i.e. using the masculine form of a term even when referring to mixed-gender groups, which erases women and non-binary individuals by treating male experience as “the default” (Mucchi-Faina, 2005). To address this, Greek institutions have proposed gender-inclusive strategies, such as 1) combined forms (“ο καθηγητής / η καθηγήτρια”), 2) combined suffixes (“ο/η καθηγητής/τρια”), and 3) exclusive feminine forms for female entities. However, these approaches assume a binary gender framework, implying that references to the feminine and masculine are supposedly exhaustive (Ntouvlis, 2020). Unlike English, which has adopted the singular ‘they’, Greek language reform has progressed at a slower pace, and equivalent options are not available, yet.

The most accessible inclusive mechanism in Greek today, also included in the aforementioned guides, is **gender-neutral language** and can be achieved through techniques like passive syntax, second-person plural, imperatives, circumlocution, and neuter grammatical gender (Kalfadopoulou and Tsigou, 2022). Each technique has its limitations and is context-dependent, making application particularly challenging. For our study on occupational nouns, circumlocution and neuter forms are the most viable techniques. As noted by Piergentili et al. (2023), gender-neutral rephrasings and synonyms is a workable paradigm toward more inclusive MT when gender is unknown or simply irrelevant. For example, gendered “καθηγητές/καθηγήτριες” (professors [M/F]) can be replaced with neutral expressions like “το διδακ-

τικό προσωπικό” (the teaching staff) or “τα μέλη του διδακτικού προσωπικού” (the members [N] of the teaching staff). In this way, we restructure sentences to eliminate gendered language, adopting neuter terms like “το άτομο” (the individual) or “το μέλος” (the member). Such neuter forms are also used by non-binary individuals, alongside neologisms like “το φίλο” (the friend), which, as any newly coined word, is rather far from earning mainstream acceptance. Finally, with regard to written discourse, gender-neutral symbols like @ (used as a suffix, e.g. “τ@ φίλ@”) are increasingly becoming popular on social media (Ntouvlis, 2020).

2.3 Gender bias in MT & LLMs as gender rewriters

In MT, we document previous research on gender bias focused on coreference resolution and pronoun translation in relation to human entities (Rudinger et al., 2018; Zhao et al., 2018; Prates et al., 2019; Cho et al., 2019; Stanovsky et al., 2019; Kocmi et al., 2020; Gonen and Webster, 2020; Levy et al., 2021; Currey et al., 2022; Robinson et al., 2024). The analyses show that popular MT systems are significantly prone to perpetuate but also exacerbate biases through systematic gender-related translation errors, while underlining the challenges of gender bias mitigation.

Approaches to this problem have involved training models from scratch on artificially gender-balanced datasets (Zhao et al., 2018; Zmigrod et al., 2019), using debiased embeddings (Bolukbasi et al., 2016; Escudé Font and Costa-jussà, 2019), and annotating data with speakers’ gender information (Vanmassenhove et al., 2018). Additional methods include POS tagging (Elaraby et al., 2018), word-level gender tagging (Stafanovičs et al., 2020; Saunders and Byrne, 2020), fine-tuning (Saunders et al., 2020), or gender re-inflection of references such as Google Translate (Johnson, 2020) and Fairslator⁴ (Měchura, 2022). However, as Savoldi et al. (2021) highlight, there is no definitive, state-of-the-art solution for mitigating bias in MT; instead, these interventions typically address isolated aspects of the problem with targeted, modular solutions. It is also worth noting that most of these studies largely operate within a binary framework, emphasising masculine and feminine forms into grammatical languages, which ultimately lim-

⁴<https://www.fairslator.com/>

its their inclusivity.

With the advent of artificial intelligence (AI), the translation capabilities and potential of LLMs in bias mitigation are being explored. Ghosh and Caliskan (2023) showed that GPT-3 tends to reinforce stereotypes and struggles with gender-neutral pronouns, as it systematically converts them into binary forms between English and Bengali, as well as five other low-resource languages (Farsi, Malay, Tagalog, Thai, Turkish). In a related study, Vanmassenhove (2024) observed a strong male bias in GPT-3.5’s English-Italian translation, despite being explicitly instructed to provide all possible gender alternatives. Lee et al. (2024) tested prompting GPT-3.5 Turbo and Llama 2 70b Chat⁵ for English to Spanish, French, and Italian translations, and despite showing promise for controlled gendered outputs, their approach remained limited to binary representations.

Another study by Sánchez et al. (2024) tested few-shot prompting on Llama-7B⁶ and showed a sufficient level of control over binary gender forms in 25 languages, underlining that similar strategies could be applicable to non-binary alternatives. Piergentili et al. (2024) extended this approach using the Neo-GATE dataset that incorporates non-binary structures and found that GPT-4 and Mixtral⁷ performed best with few-shot prompting for English-Italian translations.

LLMs like GPT-4 have shown potential for gender-neutral translations when prompted with few-shot exemplars, though zero-shot performance remains inadequate (Savoldi et al., 2024). Rarrick et al. (2024) developed a translation-rewriting solution with GPT-4, using chain-of-thought prompting, which involved explicitly providing the LLM with step-by-step reasoning and detailed clarifications in the examples. The results indicate that, while the model achieved high accuracy in pronoun adjustments, it struggled with gendered nouns, showing a limitation in handling complex gender adjustments.

3 Methodology

3.1 Dataset Preparation

We created GendEL, a manually crafted dataset of 240 gender-ambiguous and gender-unambiguous

⁵<https://huggingface.co/meta-llama/Llama-2-70b-chat-hf>

⁶<https://huggingface.co/meta-llama/Llama-2-7b>

⁷https://huggingface.co/docs/transformers/en/model_doc/mixtral

English sentences, based on a list of 40 occupational nouns. For each occupation, we created a subset of six sentences: a baseline (‘ambiguous base’) and five variations, which modify the baseline in a specific way (e.g. by adding a stereotypical adjective or a pronoun). The baseline template is: *The [OCCUPATION] finished the work.*⁸

The occupational nouns were selected from Troles and Schmid (2021), using data from the US Bureau of Labor Statistics (2019). Occupations were classified as male- or female-biased if over 50% of workers were men or women, respectively; e.g. 93% of carpenters are men, 80% of librarians are women (Appendix A). We created 20 subsets for male- and 20 for female-biased occupations.

To enrich the scope of the study, five additional sentence types were derived from the ‘ambiguous base’, resulting in six types overall (examples are provided in Table 1):

- **Ambiguous + male-biased adj.:** Ambiguous sentence with male-biased adjective.
- **Ambiguous + female-biased adj.:** Ambiguous sentence with female-biased adjective.
- **Unambiguous [Male]:** Unambiguous sentence with a masculine pronoun.
- **Unambiguous [Female]:** Unambiguous sentence with a feminine pronoun.
- **Ambiguous / unambiguous [Non-binary]:** Uses singular ‘they’, which makes the sentence either gender-ambiguous (gender is purposefully omitted or undefined) or unambiguous (referring to a non-binary individual).

The gender-biased adjectives were sourced from Troles and Schmid (2021), selecting 10 male- and 10 female-biased adjectives that we evenly and randomly distributed across the dataset (Appendix A). These sentences were kept short and simple, minimising linguistic diversity. This ensured that there were no factors influencing the referents’ gender other than the investigated words, i.e. occupational nouns, pronouns, and adjectives.

All sentences were manually translated into Greek by the author of this study. Ambiguous sentences received three translations (masculine, feminine, neutral), while unambiguous ones had

⁸Inspired by Saunders and Byrne (2020), who used binary-gendered examples like “The actor finished her/his work” for gender debiasing in translation.

Sentence Type	Example
ambiguous base	The assistant finished the work.
ambiguous + male-biased adj.	The <i>eminent</i> assistant finished the work.
ambiguous + female-biased adj.	The <i>sassy</i> assistant finished the work.
unambiguous [Male]	The assistant finished <i>his</i> work.
unambiguous [Female]	The assistant finished <i>her</i> work.
ambiguous / unambiguous [Non-binary]	The assistant finished <i>their</i> work.

Table 1: Examples of sentence types representing a subset (for occupational noun “assistant”) from GendEL.

a single correct translation. For the ‘ambiguous / unambiguous [Non-binary]’, we consider that there is only one correct translation which is a gender-neutral one. A sample of GendEL, including male- and female-biased subsets, is in [Appendix B](#).

3.2 Translation Systems

We first test two widely used commercial MT models: Google Translate and DeepL. Both of these systems have implemented a feature that provides two outputs for short gender-ambiguous queries. However, while Google Translate offers this feature for some languages, Greek is not among the supported languages for gender-ambiguous sentence outputs. On the other hand, DeepL provides this feature for Greek, but its implementation is inconsistent across different sentence structures and contexts. All sentences of GendEL were translated with both MT systems. For gender-ambiguous sentences we reported the main output and any gender alternatives provided by DeepL.

In addition to these models, we included GPT-4o into our analysis to explore its potential for gender-inclusive translations. The advantage of using an LLM is that it can be directed using a prompt, allowing for customised outputs. To achieve this, we leveraged OpenAI’s GPT-4o⁹ via the OpenAI API and, similarly to the approach of [Rarrick et al. \(2024\)](#), we used few-shot chain-of-thought prompting to encourage the model to produce gender rewrites when necessary, i.e. in gender-ambiguous cases. The full prompt is available in [Appendix D](#).

3.3 Annotation

For the finalisation of the dataset, we manually annotated the translations generated by the two MT systems and LLM with labels to indicate the gender representation in the output. Particularly, the labels included **M** (masculine), **F** (feminine), **N** (neutral),

or combinations thereof, such as **M-F-N**, **M-F**, and **M-N**, to capture cases with alternative translations. Additionally, we introduced four distinct error labels to classify certain issues:

- **error [1]:** Incorrect or nonsensical translation (e.g., non-existing words, singular replaced with plural etc.).
- **error [2]:** Mixed genders in translation (e.g., masculine noun with feminine pronoun: “The *male* farmer finished *her* work.”).
- **error [3]:** Erroneous or insufficient neutralisation techniques (e.g., using double forms or epicene nouns which imply binary gender and do not reflect true gender neutrality).
- **error [4]:** Adjective missing from the translation.

3.4 Evaluation

The evaluation of the models’ outputs was performed using a mixed-methods approach. Automatic evaluation metrics such as BLEU ([Papineni et al., 2002](#)) or TER ([Snover et al., 2006](#)) will not be included due to their known limitations. These metrics, while commonly used to evaluate translation accuracy, treat all errors equally and lack sensitivity to certain linguistic phenomena, such as gender bias ([Sennrich, 2017](#)). Following [Freitag et al. \(2021\)](#), we acknowledge that human evaluation remains the gold standard for evaluating translation quality, and we therefore rely on it to assess the translations’ validity and alignment with gender-inclusive practices in Greek.

3.4.1 Gender Bias in MT Systems

To investigate the presence of gender bias in the MT systems, we examined three key patterns of bias: *i*) **male bias**, *ii*) **occupational stereotyping**, and *iii*) **errors in anti-stereotypical gender assignments**.

⁹Model version: gpt-4o-2024-08-06

Male Bias We tested whether Google Translate and DeepL exhibit a tendency to default to masculine forms when translating gender-ambiguous English sentences into Greek. To evaluate this, we calculated the distribution of gendered outputs exclusively for the ambiguous sentences,¹⁰ where no explicit cues were provided in the source text. By analysing these trends, we aimed to identify systematic male bias in the systems’ translation behaviour.

Occupational Stereotyping We examined whether the MT systems reinforce traditional gender roles associated with specific professions (e.g. “male doctor”, “female nurse”), focusing again on the gender-ambiguous sentences. The frequency of stereotyping was calculated for male- and female-biased occupations, with statistical significance tested using Fischer’s exact test (Fisher, 1992).

Anti-Stereotypical Gender Assignments We analysed the outputs of gender-unambiguous English sentences¹¹ comparing error rates between anti-stereotypical (e.g. “female doctor”) and stereotypical cases (e.g. “male doctor”). Fischer’s exact test was used, followed by qualitative analysis of significant cases.

3.4.2 GPT-4 on Gender Bias Mitigation

We evaluated prompted GPT-4o’s ability to produce gender-inclusive translations. First, we calculated the gender distribution across all sentence types of GendEL. Second, we analysed the error distribution and conducted a qualitative review to identify factors influencing the model’s performance and highlighted areas where it deviated in terms of gender-inclusive practices.

4 Results

A preliminary analysis reveals substantial differences in gender and error distributions across the three models (Table 2). Google Translate and DeepL strongly favoured masculine forms (65.8% and 63.3%), with significantly fewer feminine translations (16.2% and 25.8%). In contrast, prompted GPT-4o showed a more balanced approach, generating 16.7% masculine, 15.8% feminine, and 12.9% neutral translations. For 42.9% of

¹⁰‘ambiguous base’, ‘ambiguous + male-biased adj.’, ‘ambiguous + female-biased adj.’, ‘ambiguous + unambiguous [Non-binary]’

¹¹‘unambiguous [Male]’, ‘unambiguous [Female]’

sentences it generated three gendered alternatives (M-F-N). These results align more closely with the gold standard, which aims for equal representation of gendered and neutral translations (16.7% for each gender and 50% for alternatives). The gold standard highlights the gap between observed and ideal distributions, particularly the lack of gender-neutral forms and alternatives by Google Translate and DeepL.

Regarding errors, Google Translate had the highest rates: 13.3% of translations included mixed genders, 3.7% were incorrect/nonsensical translations, and 0.4% omitted the adjective. DeepL performed better, with only 1.3% mixed-gender and 1.3% incorrect/nonsensical translations. GPT-4o also showed low error rates (1.7% mixed-gender, 1.3% incorrect/nonsensical) but omitted adjectives in 4.2% of cases. Interestingly, GPT-4o introduced a unique error type; as the only model that actively attempted to provide neutral forms, it did not always succeed, resulting in 3.3% of cases that contained errors related to the neutralisation techniques.

4.1 Gender Bias in MT Systems

Male Bias An analysis of the 160 ambiguous English sentences revealed a clear tendency towards male bias. Masculine forms dominated, making up 74.4% of Google Translate’s and 70.6% of DeepL’s outputs. Feminine forms were rare (8.1% for Google Translate, 13.1% for DeepL), and errors were higher for Google Translate (17.5% vs. DeepL’s 3.7%) (Appendix D).

Most errors by Google Translate were mixed-gender representations all of which appeared in sentences with the singular ‘they’, indicating a difficulty in handling the gender neutrality or non-binarity expressed by the pronoun. Specifically, the model treated ‘they’ as a collective pronoun, defaulting to masculine forms for professions and collective ‘they’ for the pronoun (i.e. “their work”), failing to correlate the gender-neutral pronoun with gender-neutral solutions in Greek.¹² DeepL’s errors, albeit very few, were found in the same ‘ambiguous + unambiguous [Non-binary]’ sentence type, suggesting that DeepL also struggles with sentences involving neutral or non-binary pronouns.

Regarding gender-inclusive outputs, neither system performed well. Google Translate did not pro-

¹²E.g., “The guard finished their work” was translated into “Ο φύλακας τελείωσε τη δουλειά τους” (= *The guard [M] finished their [plural] work*).

Label	Google Translate	DeepL	Prompted GPT-4o	Gold standards
M	158 (65.8%)	152 (63.3%)	40 (16.7%)	40 (16.7%)
F	39 (16.2%)	62 (25.8%)	38 (15.8%)	40 (16.7%)
N	1 (0.4%)	-	31 (12.9%)	40 (16.7%)
M-F-N	-	-	103 (42.9%)	120 (50%)
M-F	-	19 (7.9%)	3 (1.3%)	-
M-N	-	1 (0.4%)	-	-
error [1]	9 (3.7%)	3 (1.3%)	4 (1.7%)	-
error [2]	32 (13.3%)	3 (1.3%)	3 (1.3%)	-
error [3]	-	-	8 (3.3%)	-
error [4]	1 (0.4%)	-	10 (4.2%)	-
Total	240 (100%)	240 (100%)	240 (100%)	240 (100%)

Table 2: Distribution of gender and error labels across the three systems and gold standards, with raw counts and proportions.

duce any neutral or inclusive translations, while DeepL provided alternatives for 12.5% of gender-ambiguous sentences, mostly masculine-feminine pairs, with only one case including a neutral variant. These results confirm a **notable male bias** in both systems, aligning with common findings about male default bias in MT systems.

Occupational Stereotyping Masculine translations prevailed in both systems, regardless of stereotype: 82.5% (Google Translate) and 86.2% (DeepL) for male-biased occupations, and 66.2% (Google Translate) and 55% (DeepL) for female-biased ones. These findings agree with those in the prior section regarding the persistence of male bias in MT systems.

A closer examination of the results reveals an interesting pattern in feminine gender outputs. In the case of male-biased occupations, none of the MT systems produced translations in the feminine form. On the contrary, when the occupation was female-biased, Google Translate generated 16.2% and DeepL 26.2% feminine outputs. This signifies that, while masculine remains the default, MT systems are potentially influenced by societal stereotypes, associating feminine forms more frequently with traditionally female-biased professions.

A Fischer’s exact test confirmed a significant correlation between occupational stereotypes and gender outputs. The test results for both systems (Appendix G) verified that the stereotype of the occupation significantly impacts the translation gender, with **feminine forms more likely to appear for stereotypically female occupations** than stereotypically male ones. Although stereotypes influ-

enced translations, the masculine form remained the overall default.

Anti-Stereotypical Gender Assignments Regarding this pattern of bias, the MT systems were expected to generate more frequent errors or incorrect gender assignments when translating anti-stereotypical gender roles, such as “female doctors” or “male nurses”, compared to stereotypical ones, like “male doctors” or “female nurses”. For this pattern of bias, only sentences with feminine and masculine genders were analysed (Appendix F).

An examination of the **stereotypical cases** revealed that for male-biased occupations, both systems correctly translated masculine-gendered sentences (‘unambiguous [Male]’), where the gender ambiguity was resolved with the use of a masculine pronoun. Only one lexical error by Google Translate was detected: the word “mover” was translated as “μετακινούμενος”, which refers to someone being “moved” rather than the profession of a “mover”.

For female-biased occupations in female-gendered sentences (‘unambiguous [Female]’), DeepL demonstrated consistently accurate performance, correctly assigning the feminine gender to all translations. Google Translate, however, produced three errors (two mixed-gender, one lexical error). Specifically, the system introduced a non-existent feminised form (“φούρνισαρη”) as the translation of “female baker”. This potentially indicates not only a grammatical error but also a difficulty in handling feminine forms for some professions.

Regarding the **anti-stereotypical cases**, our initial claim is not supported for ‘unambiguous

[Male]’ sentences containing female-biased occupations,¹³ as Google Translate produced accurate masculine translations in all cases. DeepL also performed well, with only one incorrect gender assignment: “housekeeper” was translated into the feminine form.

For ‘unambiguous [Female]’ sentences containing male-biased professions,¹⁴ DeepL again showed strong performance, translating all instances into feminine forms. However, Google Translate exhibited notable variability: only 45% of the sentences were correctly assigned a feminine gender, 50% were classified as errors (mostly mixed-gender representations), and one instance was labelled as ‘other’¹⁵ (translating “female guard” as “η φρουρά”; neutral term used to refer to the role without specifying the gender). A notable error was translating “construction worker” as “η οικοδομή” (the building; lexical error), a noun with a feminine grammatical gender, suggesting that the model, in an attempt to assign feminine form to the output, used an incorrect – yet feminine – term.

Fischer’s exact test (Appendix G) revealed a statistically significant difference for **Google Translate** in feminine-gendered sentences, showing it **struggles more when translating male-biased professions into feminine forms**, despite the presence of an explicit feminine pronoun in the source. DeepL outperformed Google Translate, consistently assigning correct genders regardless of stereotypicality. Therefore, the issues in the processing of anti-stereotypical gender assignments remain specific to Google Translate.

4.2 GPT-4 on Gender Bias Mitigation

4.2.1 Quantitative Analysis

The performance of the prompted GPT-4o showed promising results overall. Appendix H presents the gender distribution across sentence types. For the ‘unambiguous [Male]’ sentences, the model achieved a 100% success rate, correctly translating all instances into masculine forms. For ‘unambiguous [Female]’ sentences, it succeeded in 95% of cases, with 5% errors.

Furthermore, for the ‘ambiguous base’, ‘ambiguous + male-biased adj.’, and ‘ambiguous +

female-biased adj.’ sentences, the model showed high precision in detecting gender ambiguity, and generated three alternatives (M-F-N) with success rates of 92.5%, 80% and 85%, respectively. The remaining cases were classified as errors.

Finally, for the ‘ambiguous / unambiguous [Non-binary]’ sentences, GPT-4o successfully produced neutral translations using the neutral circumlocution “το άτομο που” (the person who). However, 7.5% of translations in this category included double forms (e.g. “ο/η λογιστής/λογίστρια” – the [M/F] accountant [M/F]), despite explicit instructions to avoid them. 15% of translations were classified as errors.

The above results show that the masculine-gendered sentences obtained the highest accuracy, followed by the feminine-gendered ones. This indicates that the model performed best when the gender of the referent was explicitly specified, with a slightly reduced success rate for feminine forms, possibly reflecting intrinsic bias present in the training data.

A closer comparison of the performance of prompted GPT-4o with Google Translate and DeepL on unambiguous sentences reveals that the LLM outperformed the other two MT systems with 100% success rate in masculine-gendered sentences (Appendix I). Interestingly, in **feminine-gendered sentences**, it was DeepL that had the highest scores (100%), followed closely by GPT-4o (95%). Google Translate, in contrast, achieved only 65% accuracy, with a notable number of incorrect gender assignments and errors. GPT-4o’s slight reduction in accuracy for feminine forms may suggest residual biases in its training data.

Overall, GPT-4o handled gender ambiguity effectively, generating three correct alternatives in most cases. However, when (gender-biased) adjectives were included, the rates slightly dropped, indicating that such modifiers introduce additional difficulty. The ‘ambiguous / unambiguous [Non-binary]’ sentence type posed the greatest challenge for the model, with the highest percentage of incorrect translations, which may reflect the limitations of the model in producing accurate gender-neutral language.

4.2.2 Qualitative Analysis

A closer investigation of the errors for each sentence category shows interesting information regarding the inaccuracies produced by prompted GPT-4o (Appendix H). Key observations include:

¹³E.g., “The teacher [F-biased] finished *his* work.”

¹⁴E.g., “The driver [M-biased] finished *her* work.”

¹⁵Outputs that included masculine-feminine alternatives (M-F), masculine-neutral alternatives (M-N), or exclusively gender-neutral forms (N) were grouped together under the ‘other’ category for illustration purposes.

Ambiguous base The model consistently produced three gender alternatives for each source sentence, in which the masculine and feminine versions were accurate. However, the neutral forms presented issues: the model attempted to use the neuter circumlocution “το άτομο που” (the person who) but paired with epicene nouns, such as “μηχανικός” (mechanic), terms that maintain binary gender distinctions and fail to accommodate non-binary references. This demonstrates both the model’s challenges in achieving true gender neutrality, but also the broader constraints of Greek’s available neutralisation strategies.

Ambiguous + male-biased adj. Once again, all errors occurred in the gender-neutral variants provided by the model, alongside the correct masculine and feminine versions. Most issues in this category, involved the omission of the male-biased adjective from the target sentence. This pattern suggests that the model may struggle to balance its neutralisation efforts with preserving the semantic elements of the source text. In other words, we assume that the model prioritised neutralisation to such an extent that it overlooked key details, such as the male-biased adjective, which is critical for maintaining the original meaning of the sentence.

Ambiguous + female-biased adj. The model’s neutral variants again exhibited three issues: (a) omission of the female-biased adjective in three cases, compromising the original meaning, (b) inaccurate neutralisation attempts using epicenes in two instances, and (c) one syntactically incorrect output featuring redundant repetition of “που” ([the person] who).

Unambiguous [Female] The errors here were minimal, with two instances classified as mixed-gender representations. This error shows a mismatch between the grammatical gender of the subject “sheriff” and “farmer” and the personal pronoun ‘her’ (translated as “The *male* sheriff/farmer finished *her* work”). While these translations technically align with the source text in terms of pronoun use, the introduction of a masculine article and noun creates a bias and inconsistency that makes the translation somewhat problematic. Instead of fully aligning the output gender-wise based on the feminine pronoun, it defaulted to the masculine form of “ο σερίφης” (the male sheriff) and “ο αγρότης” (the male farmer), possibly influenced by inherent bias of the training data.

Ambiguous / unambiguous [Non-binary] GPT-4o’s most significant challenges emerged in this category, revealing difficulties in producing sufficient gender-neutral translations. First, it frequently defaulted to inadequate solutions, either employing epicenes, such as “μηχανικός” (mechanic) and “υπάλληλος” (clerk), or binary double forms, such as “ο/η συντάκτης/τρια” (the [M/F] editor [M/F]) incorrectly paired with the plural pronoun “τους” (their). This indicates that the model misinterpreted singular ‘their’ as a collective pronoun.

Second, the model generated linguistically invalid forms while attempting neutralisation, including (i) the non-existent “το ρεσεψιονίστ”, (ii) repeated feminine possessives in “τη δουλειά του/της/της” (his/her/her work), (iii) the ill-formed “ο/η/το γραμματέας” (the [M/F/N] secretary), suggesting a non-existent neuter form of the specific epicene, and (iv) the completely invented pronoun “ατους”. The model apparently struggled to produce gender-neutral language leading to mistranslations and non-existent words.

Third, rather than producing a singular gender-neutral form, GPT-4o often defaulted to listing multiple gendered variations, contradicting prompt instructions for singular ‘they’ translations. In three cases, it generated grammatically correct but non-inclusive outputs, failing to fully adhere to non-binary representations.

5 Discussion

Our study confirms significant gender bias in English-to-Greek MT, with both Google Translate and DeepL defaulting to masculine forms for gender-ambiguous contexts, rendering it consistent with findings in other language pairs (Prates et al., 2019; Stanovsky et al., 2019; Currey et al., 2022). While this male bias persisted across occupational stereotypes, increased feminine forms for stereotypically female occupations demonstrate how systems simultaneously reinforce male defaults and societal gender associations embedded in the systems’ training data (Savoldi et al., 2021).

Both MT systems performed consistently well with explicitly masculine referents, supporting previous work showing better handling of male references and stereotypical roles (Sun et al., 2019; Kocmi et al., 2020; Saunders and Byrne, 2020). However, Google Translate struggled more with anti-stereotypical feminine forms. This disparity is likely influenced by differences in training data

or model architecture, but the black-box nature of these systems makes it difficult to determine the exact cause. Most critically, neither system produced (adequate) gender-neutral outputs, consistently failing on non-binary cases.

Prompted GPT-4o demonstrated high performance, successfully generating feminine, neutral and masculine alternatives for most ambiguous cases. While it showed residual bias (better accuracy in masculine forms), the model was able to follow instructions to identify gender-ambiguous sentences and generate inclusive output for them. We thus confirmed LLMs' potential to handle gender-aware translation, supporting emerging research (Savoldi et al., 2024) that also demonstrates that GPT is a promising solution for producing gender-neutral outputs when given only a few examples.

Nevertheless, the translation errors (e.g. non-existent words, incorrect pronouns, missing adjectives) or incorrect neutralisation solutions (e.g. use of epicene nouns, double forms with binary pronouns) that were produced for a small number of gender-neutral outputs, should be taken into consideration. These issues underscore the tension between technological solutions and linguistic reality; that is, an inherent bias in the training data of the model, but also the challenges in adapting gender-neutral practices for Greek. Language reform in Greek has followed a slower pace compared to other languages and currently lacks sufficient linguistic structures for a gender-neutral language or structures that address the visibility of under-represented groups, such as LGBTQIA+ individuals and women. As such, the discussion about linguistic sexism and the development of gender-inclusive practices is still open and evolving.

6 Limitations

Some limitations should be acknowledged. First, there is a reproducibility problem as the study relies on three closed-source models. As proprietary systems subject to frequent updates, the results of the same query may vary across multiple trials. Moreover, GPT-4o requires a paid subscription,¹⁶ which limits its accessibility compared to freely available systems.

Second, while GendEL's controlled sentence structures (based on gender-biased adjectives and occupational nouns) ensure methodological consistency, this design restricts the generalisability of the

¹⁶\$2.50/1M input tokens and \$10.00/1M output tokens

results to more diverse and natural text. Similarly, our GPT-4o prompt was specifically customised for these sentence structures, which raises questions about its applicability to more complex data. GendEL, therefore, should be viewed as a foundational resource for evaluating English-to-Greek gender bias, and future research could supplement a wider variety of sentence structures, contexts, linguistic phenomena, and manifestations of gender bias, as well as further experimentation with LLM prompting strategies.

7 Conclusion

In response to the emerging demand for inclusive language, this study focused on the under-represented English-to-Greek language pair. Through extensive, fine-grained manual analyses and descriptive statistics, we demonstrated that gender bias is persistent in translations by Google Translate and DeepL, highlighting that, while they perform well in cases where the referent's gender is defined, they are far from recognising and producing gender-neutral language. We also demonstrated that GPT-4o, when prompted, can achieve high accuracy on providing gendered and gender-neutral alternatives in cases of ambiguity. By situating our results within the context of prior research, this study makes two important contributions: (1) the creation and public release of GendEL, the first handcrafted dataset for evaluating English-to-Greek translations, and (2) empirical evidence emphasising the urgent need for more inclusive translation practices in Greek. We hope this work will inspire further research on this language pair and contribute to the development of more inclusive translation technologies.

References

- Angeliki Alvanoudi. 2015. *Grammatical gender in interaction: Cultural and cognitive aspects*. Brill.
- Omri Asscher and Ella Glikson. 2023. [Human evaluations of machine translation in an ethically charged situation](#). *New Media & Society*, 25(5):1087–1107. Publisher: SAGE Publications.
- 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 home-maker? debiasing word embeddings](#). *Preprint*, arXiv:1607.06520.
- Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. [On measuring gender bias in translation of gender-neutral pronouns](#). In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.
- 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.
- Mostafa Elaraby, Ahmed Y. Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. [Gender aware spoken language translation applied to english-arabic](#). In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6.
- Joel Escudé Font and Marta R. Costa-jussà. 2019. [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, Florence, Italy. Association for Computational Linguistics.
- R. A. Fisher. 1992. *Statistical Methods for Research Workers*, pages 66–70. Springer New York, New York, NY.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474. Place: Cambridge, MA Publisher: MIT Press.
- Batya Friedman and Helen Nissenbaum. 1996. [Bias in computer systems](#). *ACM Trans. Inf. Syst.*, 14(3):330–347.
- 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*, AIES '23, page 901–912, New York, NY, USA. Association for Computing Machinery.
- Hila Gonen and Kellie Webster. 2020. [Automatically identifying gender issues in machine translation using perturbations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics.
- Melvin Johnson. 2020. [A scalable approach to reducing gender bias in google translate](#).
- Valentini Kalfadopoulos and Maria Tsigou. 2022. [Inclusive language in translation technology: Theory and practice; the case of greek](#). In *Proceedings of the New Trends in Translation and Technology Conference - NeTTT 2022*, pages 206–213, Rhodes Island, Greece.
- Anestis Polychronis Karastergiou and Konstantinos Diamantopoulos. 2024. [Gender issues in machine translation](#). *Transcultural Journal of Humanities Social Sciences*, 5:48–64.
- Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. [Gender coreference and bias evaluation at WMT 2020](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.
- 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.
- Manuel Lardelli and Dagmar Gromann. 2023. [Translating non-binary coming-out reports: Gender-fair language strategies and use in news articles](#). *The Journal of Specialised Translation*, pages 213–240.
- Minwoo Lee, Hyukhun Koh, Minsung Kim, and Kyomin Jung. 2024. [Fine-grained gender control in machine translation with large language models](#). In *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, Mexico City, Mexico. Association for Computational Linguistics.
- Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. [Collecting a large-scale gender bias dataset for coreference resolution and machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Eva Tsouparopoulou, Dimitris Parsanoglou, Maria Symeonaki, and Giorgos Stamou. 2024. [Gostmt: A knowledge graph for occupation-related gender biases in machine translation](#). *Preprint*, arXiv:2409.10989.
- Michal Měchura. 2022. [A taxonomy of bias-causing ambiguities in machine translation](#). In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173, Seattle, Washington. Association for Computational Linguistics.

- Johanna Monti. 2020. Gender issues in machine translation: An unsolved problem? In *The Routledge Handbook of Translation, Feminism and Gender*. Routledge.
- Angelica Mucchi-Faina. 2005. [Visible or influential? Language reforms and gender \(in\)equality](#). *Social Science Information*, 44(1):189–215.
- Vinicio Ntouvli. 2020. Online writing and linguistic sexism: The use of gender-inclusive @ on a greek feminist facebook page. *Tilburg Papers in Culture Studies*, 245.
- Mary Nurminen and Maarit Koponen. 2020. [Machine translation and fair access to information](#). *Translation Spaces*, 9(1):150–169. Publisher: John Benjamins Publishing Company.
- 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.
- Theodossia-Soula Pavlidou, Angeliki Alvanoudi, and Eleni Karafoti. 2004. Grammatical gender and semantic content: preliminary remarks on the lexical representation of social gender [in Greek]. In *Studies in Greek Linguistics*, volume 24, pages 543–553. Aristotle University of Thessaloniki.
- 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).
- Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2019. [Assessing gender bias in machine translation – a case study with google translate](#). *Preprint*, arXiv:1809.02208.
- Spencer Rarrick, Ranjita Naik, Sundar Poudel, and Vishal Chowdhary. 2024. [Gate x-e : A challenge set for gender-fair translations from weakly-gendered languages](#). *Preprint*, arXiv:2402.14277.
- Kevin Robinson, Sneha Kudugunta, Romina Stella, Sunipa Dev, and Jasmijn Bastings. 2024. [Mittens: A dataset for evaluating gender mistranslation](#). *Preprint*, arXiv:2401.06935.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [Gender bias in coreference resolution](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- 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.
- Danielle Saunders and Bill Byrne. 2020. [Reducing gender bias in neural machine translation as a domain adaptation problem](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.
- Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. [Neural machine translation doesn’t translate gender coreference right unless you make it](#). In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.
- Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. [Gender bias in machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:845–874.
- Beatrice Savoldi, Andrea 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)*, pages 256–267, St. Julian’s, Malta. Association for Computational Linguistics.
- 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.
- 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.
- Artūrs Stefanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. [Mitigating gender bias in machine translation with target gender annotations](#). In *Proceedings of*

- the *Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.
- Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. [Evaluating gender bias in machine translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.
- 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.
- Μαρία Γκασούκα and Μαριάνθη Γεωργαλίδου. 2018. Οδηγός μη σεξιστικής γλώσσας στα διοικητικά έγγραφα.
- Ντενίζ Σαρρή-Χασάν. 2024. Πρακτικός Οδηγός για τη Χρήση Συμπεριληπτικής ως προς το Φύλο Γλώσσας στο ΕΑΠ. Εκδόσεις ΕΑΠ Α.Ε.
- Μανόλης Τριανταφυλλίδης. 1963. Η βουλευτίνα και ο σχηματισμός των θηλυκών επαγγελματικών, volume B, pages 326–334. Τόμος Μανόλη Τριανταφυλλίδη.
- Ρούλα Τσοκαλίδου. 1996. Το Φύλο της Γλώσσας, Οδηγός μη-σεξιστικής γλώσσας για τον δημόσιο ελληνικό λόγο. Σύνδεσμος Ελληνίδων Επιστημόνων-Βιβλιοπωλείο της Εστίας.
- Αγαπητός Τσοπανάκης. 1982. Ο δρόμος προς την δημοτική: Θεωρητικά, τεχνικά και γλωσσικά προβλήματα. Σχηματισμός επαγγελματικών θηλυκών. In *Ο δρόμος προς την Δημοτική (Μελέτες και άρθρα)*, pages 302–342. Εκδοτικός οίκος Αφών Κυριακίδη.
- 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*, pages 531–541, Online. Association for Computational Linguistics.
- Eva Vanmassenhove. 2024. [Gender bias in machine translation and the era of large language models](#). *ArXiv*, abs/2401.10016.
- Eva Vanmassenhove, Chris Emmerly, and Dimitar Shterionov. 2021. [NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. [Getting gender right in neural machine translation](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.
- Lucas Nunes Vieira. 2020. [Machine translation in the news: A framing analysis of the written press](#). *Translation Spaces*, 9(1):98–122. Publisher: John Benjamins.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. [Gender bias in coreference resolution: Evaluation and debiasing methods](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.
- Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. [Gender bias in large language models across multiple languages](#). *Preprint*, arXiv:2403.00277.
- Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. [Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Stereotypical Occupations and Adjectives

Male-biased occupations	carpenter (3%), construction worker (4%), laborer (4%), mechanic (4%), driver (20%), mover (18%), sheriff (18%), developer (20%), guard (22%), farmer (25%), chief (28%), lawyer (36%), janitor (37%), CEO (39%), analyst (41%), physician (41%), cook (42%), manager (43%), supervisor (44%), salesperson (48%)
Female-biased occupations	designer (54%), baker (60%), accountant (62%), auditor (62%), editor (63%), writer (63%), cashier (71%), clerk (72%), tailor (75%), attendant (76%), counselor (76%), teacher (78%), librarian (80%), assistant (85%), cleaner (89%), housekeeper (89%), receptionist (89%), nurse (90%), hairdresser (92%), secretary (93%)

Table 3: Male-biased and female-biased occupations included in GendEL. The percentage of women in the occupation in the US is displayed in brackets (Troles and Schmid, 2021).

Male-biased adjectives	grizzled, affable, jovial, suave, debonair, wiry, rascally, arrogant, shifty, eminent
Female-biased adjectives	sassy, perky, brunette, blonde, lovely, vivacious, saucy, bubbly, alluring, married

Table 4: Male-biased and female-biased adjectives included in GendEL.

B GendEL Sample

GendEL ID	Biased occupation	Stereotypical gender of occupation	Biased adjective	ENGLISH	Type	Gender	ONE CORRECT TRANSLATION	MULTIPLE CORRECT TRANSLATIONS			HT gender
								Alternative 1 [M]	Alternative 2 [F]	Alternative 3 [N]	
67	lawyer	M		The lawyer finished the work.	ambiguous base	N		Ο δικηγόρος τελείωσε τη δουλειά.	Η δικηγόρινα τελείωσε τη δουλειά.	Το άτομο που ασκεί τη δικηγορία τελείωσε τη δουλειά.	M-F-N
68	lawyer	M	shifty	The shifty lawyer finished the work.	ambiguous + male-biased adj.	N		Ο αναξίπιστος δικηγόρος τελείωσε τη δουλειά.	Η αναξίπιστη δικηγόρινα τελείωσε τη δουλειά.	Το αναξίπιστο άτομο που ασκεί τη δικηγορία τελείωσε τη δουλειά.	M-F-N
69	lawyer	M	perky	The perky lawyer finished the work.	ambiguous + female-biased adj.	N		Ο ευδιάθετος δικηγόρος τελείωσε τη δουλειά.	Η ευδιάθετη δικηγόρινα τελείωσε τη δουλειά.	Το ευδιάθετο άτομο που ασκεί τη δικηγορία τελείωσε τη δουλειά.	M-F-N
70	lawyer	M		The lawyer finished his work.	unambiguous [Male]	M	Ο δικηγόρος τελείωσε τη δουλειά του.				M
71	lawyer	M		The lawyer finished her work.	unambiguous [Female]	F	Η δικηγόρινα τελείωσε τη δουλειά της.				F
72	lawyer	M		The lawyer finished their work.	ambiguous / unambiguous [Non-binary]	N	Το άτομο που ασκεί τη δικηγορία τελείωσε τη δουλειά του.				N
133	accountant	F		The accountant finished the work.	ambiguous base	N		Ο λογιστής τελείωσε τη δουλειά.	Η λογίστρια τελείωσε τη δουλειά.	Το άτομο που εργάζεται στο λογιστήριο τελείωσε τη δουλειά.	M-F-N
134	accountant	F	eminent	The eminent accountant finished the work.	ambiguous + male-biased adj.	N		Ο επιφανής λογιστής τελείωσε τη δουλειά.	Η επιφανής λογίστρια τελείωσε τη δουλειά.	Το επιφανές άτομο που εργάζεται στο λογιστήριο τελείωσε τη δουλειά.	M-F-N
135	accountant	F	saucy	The saucy accountant finished the work.	ambiguous + female-biased adj.	N		Ο πονηρός λογιστής τελείωσε τη δουλειά.	Η πονηρή λογίστρια τελείωσε τη δουλειά.	Το πονηρό άτομο που εργάζεται στο λογιστήριο τελείωσε τη δουλειά.	M-F-N
136	accountant	F		The accountant finished his work.	unambiguous [Male]	M	Ο λογιστής τελείωσε τη δουλειά του.				M
137	accountant	F		The accountant finished her work.	unambiguous [Female]	F	Η λογίστρια τελείωσε τη δουλειά της.				F
138	accountant	F		The accountant finished their work.	ambiguous / unambiguous [Non-binary]	N	Το άτομο που εργάζεται στο λογιστήριο τελείωσε τη δουλειά του.				N

Figure 1: Subset of a male-biased (“lawyer”) and a subset of a female-biased (“accountant”) occupation.

C Prompt for GPT-4o

You are a machine translation assistant focused on gender-fair translations. Translate the given English text into Greek following these rules:

1. If the gender of the referent is defined, translate according to that gender. Only ONE translation is correct. E.g.:

Input: The student finished his work.

Translation: Ο μαθητής τελείωσε τη δουλειά του.

2. If the gender of the referent is not defined, provide three alternatives: masculine, feminine, neutral.
E.g.:

Input: The happy professor finished the work.

Translations:

- a) Male: Ο χαρούμενος καθηγητής τελείωσε τη δουλειά.
- b) Female: Η χαρούμενη καθηγήτρια τελείωσε τη δουλειά.
- c) Neutral: Το χαρούμενο μέλος του εκπαιδευτικού προσωπικού τελείωσε τη δουλειά.

Input: The inspector finished the work.

Translations:

- a) Male: Ο επιθεωρητής τελείωσε τη δουλειά.
- b) Female: Η επιθεωρήτρια τελείωσε τη δουλειά.
- c) Neutral: Το άτομο που επιθεωρεί τελείωσε τη δουλειά.

IMPORTANT: Pay attention to identifying the non-binary singular ‘they’ pronoun, which is used for non-binary individuals. If spotted, you must provide only the neutral version, e.g. “Το άτομο που επιθεωρεί τελείωσε τη δουλειά του”.

IMPORTANT: For the neutral version, do NOT indicate any gender. Avoid forms like “ο επιθεωρητής” or “ο/η επιθεωρητής/τρια”.

Translate this text: {input_text}

D Gender Bias in MT Systems: Male Bias

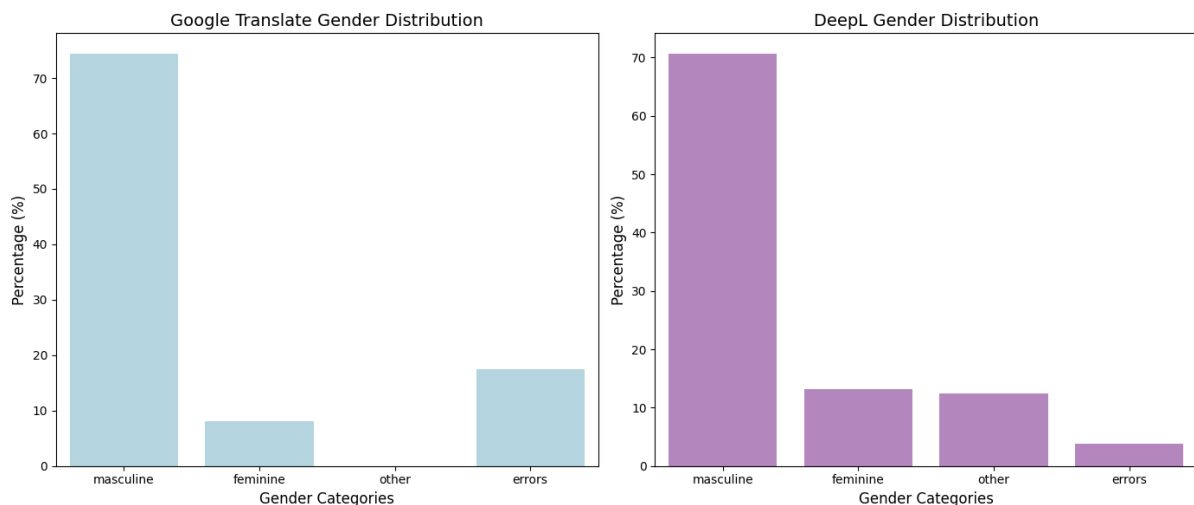


Figure 2: Gender distribution of translations for **gender-ambiguous** sentences by Google Translate and DeepL.

Gender Category	Google Translate	DeepL
masculine	119	113
feminine	13	21
other	-	20
errors	28	6
Total	160	160

Table 5: Absolute counts for gender and error categories in translations by Google Translate and DeepL.

E Gender Bias in MT Systems: Occupational Stereotyping

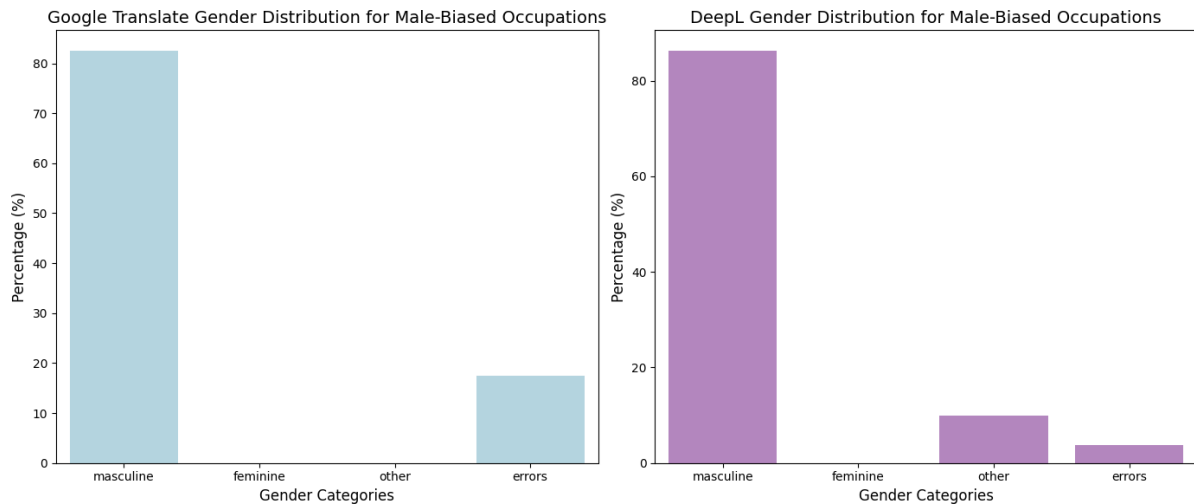


Figure 3: Gender distribution of translations for **stereotypically male occupations** in **gender-ambiguous** sentences, produced by Google Translate and DeepL.

Gender Category	Google Translate	DeepL
masculine	66	69
feminine	-	-
other	-	8
errors	14	3
Total	80	80

Table 6: Absolute counts for gender distribution of translations for **stereotypically male occupations** in **gender-ambiguous sentences**, produced by Google Translate and DeepL.

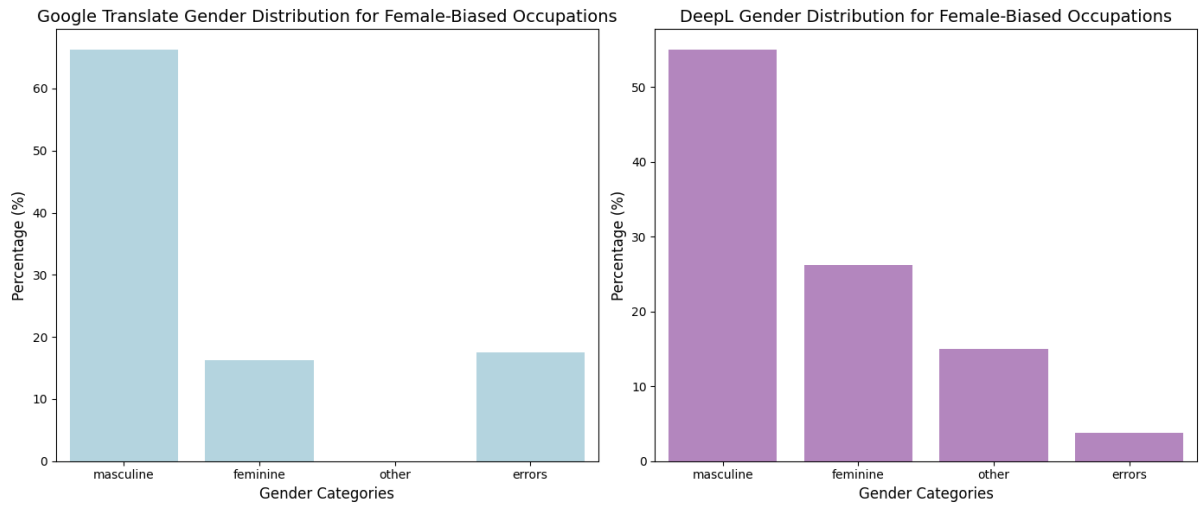


Figure 4: Gender distribution of translations for **stereotypically female occupations** in **gender-ambiguous sentences**, produced by Google Translate and DeepL.

Gender Category	Google Translate	DeepL
masculine	53	44
feminine	13	21
other	-	12
errors	14	3
Total	80	80

Table 7: Absolute counts for gender distribution of translations for **stereotypically female occupations** in **gender-ambiguous sentences**, produced by Google Translate and DeepL.

F Gender Bias in MT Systems: Anti-Stereotypical Gender Assignments

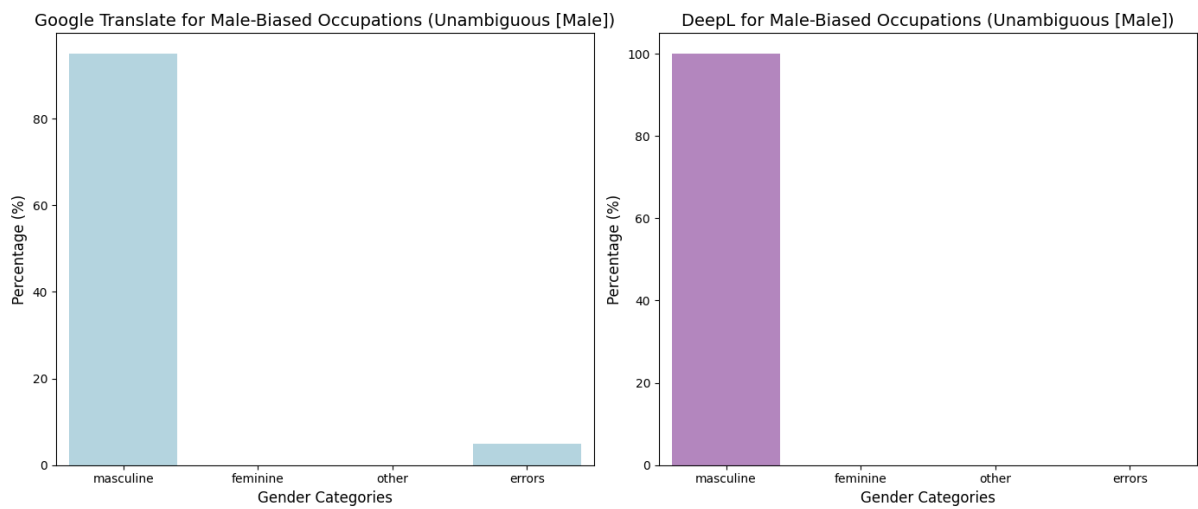


Figure 5: Gender distribution of translations for **stereotypically male occupations** in **masculine gender-unambiguous** sentences (stereotypical case), produced by Google Translate and DeepL.

Gender Category	Google Translate	DeepL
masculine	19	20
feminine	-	-
other	-	-
errors	1	-
Total	20	20

Table 8: Absolute counts for gender distribution of translations for **stereotypically male occupations** in **masculine gender-unambiguous** sentences (stereotypical case), produced by Google Translate and DeepL.

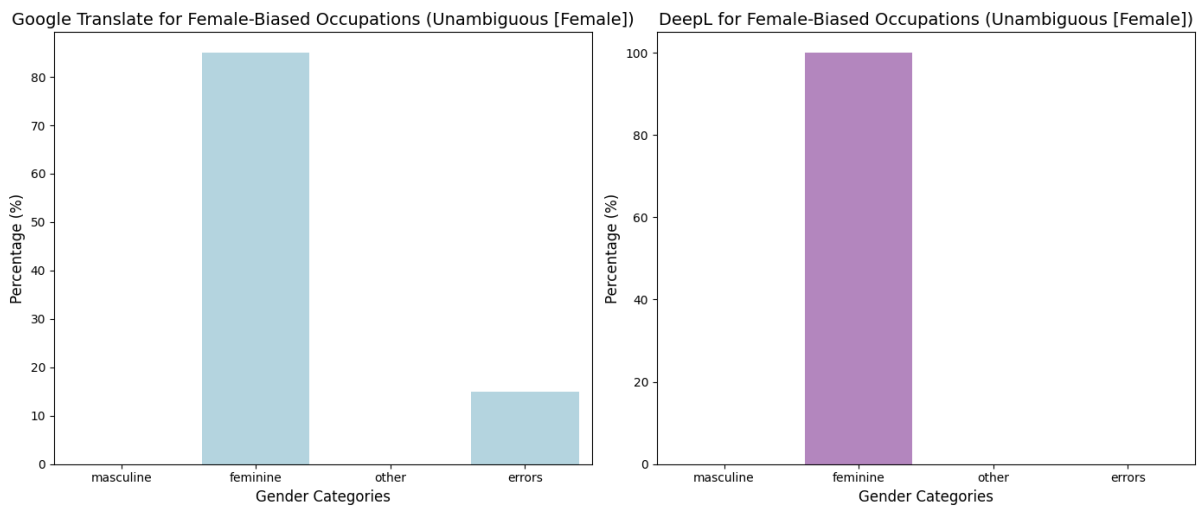


Figure 6: Gender distribution of translations for **stereotypically female occupations** in **feminine gender-unambiguous** sentences (stereotypical case), produced by Google Translate and DeepL.

Gender Category	Google Translate	DeepL
masculine	-	-
feminine	17	20
other	-	-
errors	3	-
Total	20	20

Table 9: Absolute counts for gender distribution of translations of translations for **stereotypically female occupations** in **feminine gender-unambiguous sentences** (stereotypical case), produced by Google Translate and DeepL.

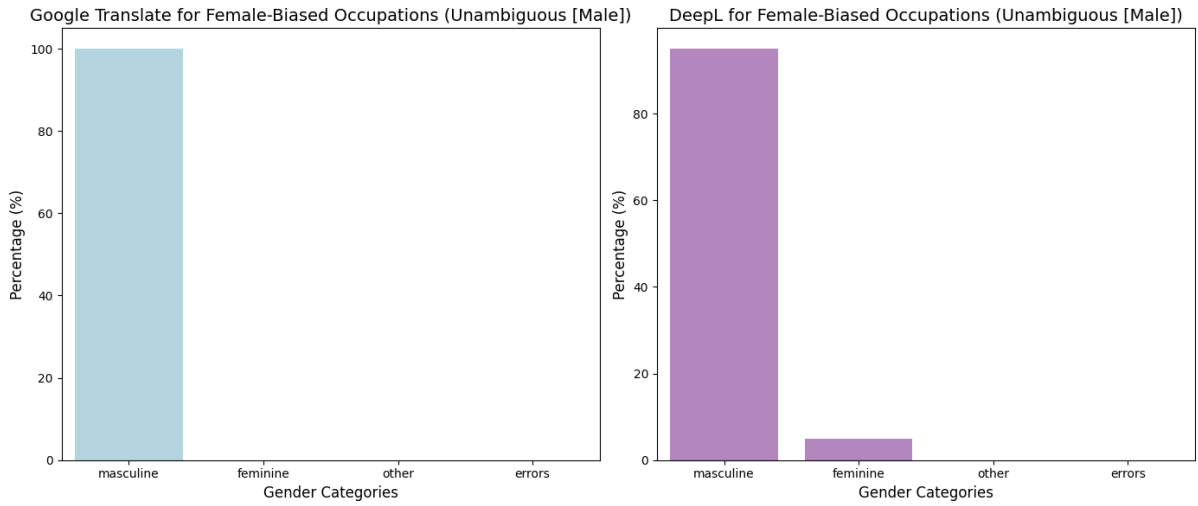


Figure 7: Gender distribution of translations for **stereotypically female occupations** in **masculine gender-unambiguous** sentences (anti-stereotypical case), produced by Google Translate and DeepL.

Gender Category	Google Translate	DeepL
masculine	20	19
feminine	-	1
other	-	-
errors	-	-
Total	20	20

Table 10: Absolute counts for gender distribution of translations for **stereotypically female occupations** in **masculine gender-unambiguous** sentences (anti-stereotypical case), produced by Google Translate and DeepL.

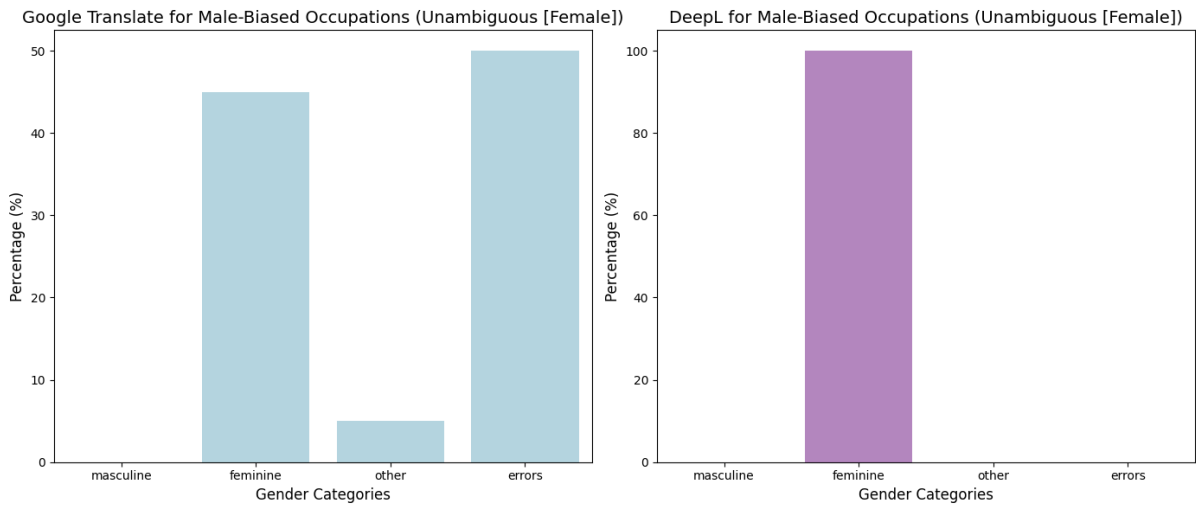


Figure 8: Gender distribution of translations for **stereotypically male occupations** in **feminine gender-unambiguous** sentences (anti-stereotypical case), produced by Google Translate and DeepL.

Gender Category	Google Translate	DeepL
masculine	-	-
feminine	9	20
other	1	-
errors	10	-
Total	20	20

Table 11: Absolute counts for gender distribution of translations for **stereotypically male occupations** in **feminine gender-unambiguous sentences** (anti-stereotypical case), produced by Google Translate and DeepL.

G Results of Fischer’s Exact Test

Metric	Google Translate	DeepL
Odds ratio	0.0	0.0
<i>p</i> -value	0.000143	2.105037e-07

Table 12: Results of Fisher’s exact test for Google Translate and DeepL, investigating whether there is a statistically significant association between the stereotype of the occupation (male- or female-biased) and the gender of the translation (e.g., masculine, feminine).

Metric	Google Translate	DeepL
Odds ratio	0.0	0.0
<i>p</i> -value	1.0	inf

Table 13: Results of Fisher’s exact test for Google Translate and DeepL, investigating whether there is a statistically significant difference between the anti-stereotypical and stereotypical groups when the gender is **masculine** and **unambiguous** in the English sentence.

Metric	Google Translate	DeepL
Odds ratio	6.925925	n/a
<i>p</i> -value	0.018701	n/a

Table 14: Results of Fisher’s exact test for Google Translate, investigating whether there is a statistically significant difference between the anti-stereotypical and stereotypical groups when the gender is **feminine** and **unambiguous** in the English sentence. For DeepL, both groups have exactly 20 feminine translations and 0 non-feminine translations, thus performing a Fisher’s exact test is not meaningful.

H GPT-4o on Bias Mitigation

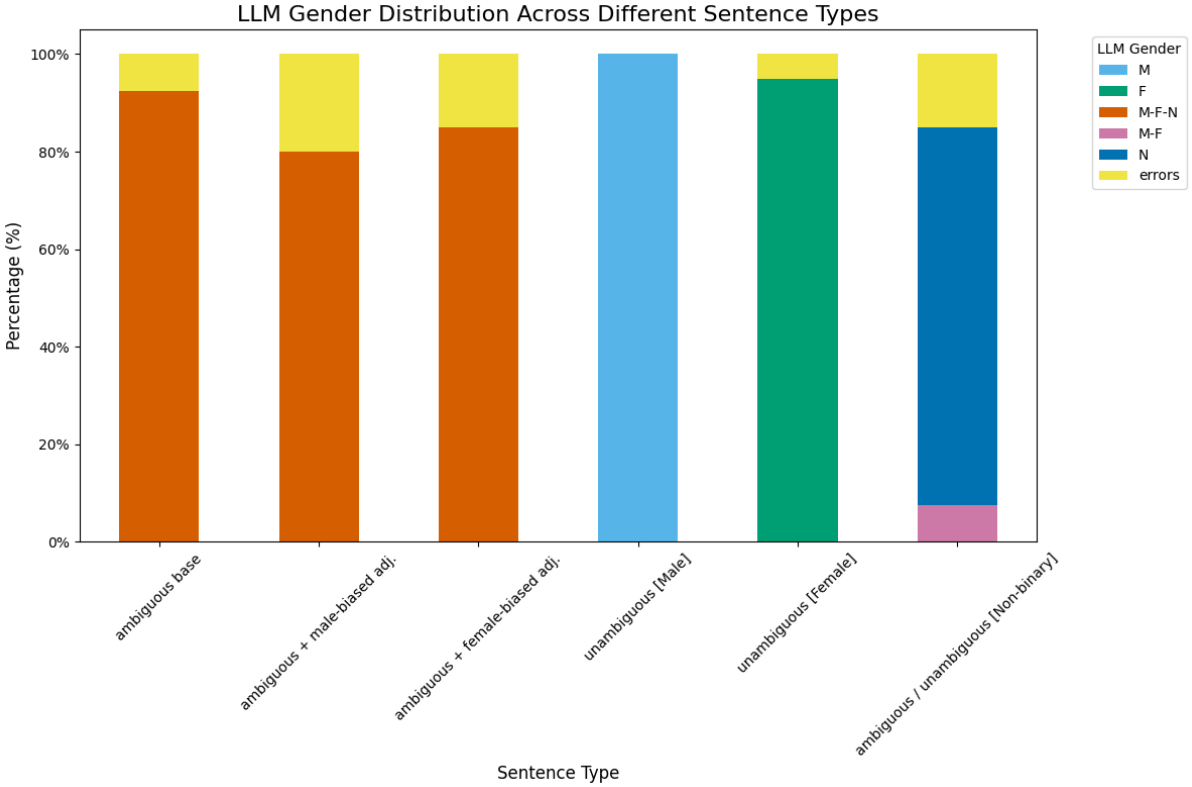


Figure 9: GPT-4o gender distribution across all sentence types.

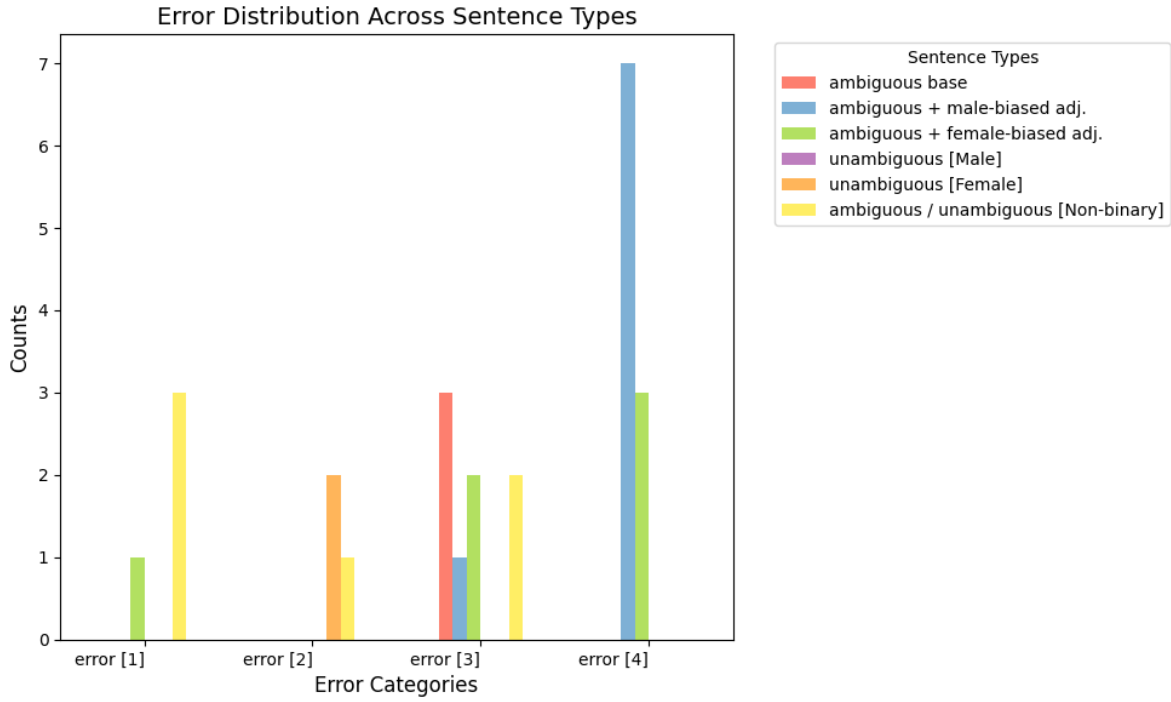


Figure 10: Error distribution of prompted GPT-4o across all sentence types.

Type	M	F	M-F-N	M-F	N	errors
ambiguous base	-	-	37 (92.5%)	-	-	3 (7.5%)
ambiguous + male-biased adj.	-	-	32 (80%)	-	-	8 (20%)
ambiguous + female-biased adj.	-	-	34 (85%)	-	-	6 (15%)
unambiguous [Male]	40 (100%)	-	-	-	-	-
unambiguous [Female]	-	38 (95%)	-	-	-	2 (5%)
ambiguous / unambiguous [Non-binary]	-	-	-	3 (7.5%)	31 (77.5%)	6 (15%)

Table 15: Absolute counts and percentages for gender categories and errors in translations by prompted GPT-4o.

Error type	English source	Translation by prompted GPT-4o
3	The mechanic finished the work.	Το άτομο που εργάζεται ως μηχανικός τελείωσε τη δουλειά. (The person who works as mechanic...)
3	The sheriff finished the work.	Το άτομο που είναι σερίφης τελείωσε τη δουλειά. (The person who is sheriff...)
3	The CEO finished the work.	Το άτομο που είναι διευθύνων σύμβουλος τελείωσε τη δουλειά. (The person who is CEO...)

Table 16: Error instances produced by prompted GPT-4o for the ‘ambiguous base’ sentences.

Error type	English source	Translation by prompted GPT-4o
3	The suave mechanic finished the work.	Το κομψό άτομο που εργάζεται ως μηχανικός τελείωσε τη δουλειά. (The suave person who works as mechanic...)
4	The grizzled mover finished the work.	Το άτομο που ασχολείται με τις μεταφορές τελείωσε τη δουλειά.
4	The arrogant sheriff finished the work.	Το άτομο της σερίφειας αρμοδιότητας τελείωσε τη δουλειά.
4	The grizzled developer finished the work.	Το άτομο που ασχολείται με τον προγραμματισμό τελείωσε τη δουλειά.
4	The jovial guard finished the work.	Το άτομο που φρουρεί τελείωσε τη δουλειά.
4	The shifty lawyer finished the work.	Το άτομο που ασκεί τη δικηγορία τελείωσε τη δουλειά.
4	The shifty writer finished the work.	Το άτομο που γράφει τελείωσε τη δουλειά.
4	The rascally tailor finished the work.	Το άτομο που ράβει τελείωσε τη δουλειά.

Table 17: Error instances produced by prompted GPT-4o for the ‘ambiguous + male-biased adj.’ sentences.

Error type	English source	Translation by prompted GPT-4o
1	The bubbly carpenter finished the work.	Το ζωηρό άτομο που που ασχολείται με την ξυλουργική τελείωσε τη δουλειά. (repetition of “που” [who])
3	The brunette mechanic finished the work.	Το άτομο με τα μελαχρινά μαλλιά που είναι μηχανικός τελείωσε τη δουλειά.
3	The vivacious CEO finished the work.	Το ζωηρό άτομο στη θέση του διευθύνοντος συμβούλου τελείωσε τη δουλειά.
4	The perky lawyer finished the work.	Το άτομο που ασκεί τη δικηγορία τελείωσε τη δουλειά.
4	The perky manager finished the work.	Το άτομο σε θέση διεύθυνσης τελείωσε τη δουλειά.
4	The alluring auditor finished the work.	Το άτομο που ελέγχει τελείωσε τη δουλειά.

Table 18: Error instances produced by prompted GPT-4o for the ‘ambiguous + female-biased adj.’ sentences.

Error type	English source	Translation by prompted GPT-4o
2	The sheriff finished her work.	Ο σερίφης τελείωσε τη δουλειά της. (The male sheriff finished her work.)
2	The farmer finished her work.	Ο αγρότης τελείωσε τη δουλειά της. (The male farmer finished her work.)

Table 19: Error instances produced by prompted GPT-4o for the ‘unambiguous [Female]’ sentences.

Error type	English source	Translation by prompted GPT-4o
1	The writer finished their work.	Η συγγραφέας/ο συγγραφέας τελείωσε τη δουλειά ατους.
1	The receptionist finished their work.	Η/Ο/Το ρεσεψιονίστ/ρεσεψιονίστρια/ρεσεψιονίστ τελείωσε τη δουλειά του/της/της.
1	The secretary finished their work.	Η/Ο/Το γραμματέας τελείωσε τη δουλειά του/της.
2	The editor finished their work.	Ο/Η συντάκτης/τρια τελείωσε τη δουλειά τους.
3	The mechanic finished their work.	Το άτομο που εργάζεται ως μηχανικός τελείωσε τη δουλειά.
3	The clerk finished their work.	Το άτομο που εργάζεται ως υπάλληλος τελείωσε τη δουλειά του.
M-F	The developer finished their work.	Ο/Η προγραμματιστής/τρια τελείωσε τη δουλειά του/της.
M-F	The farmer finished their work.	Το άτομο που δουλεύει στη φάρμα τελείωσε τη δουλειά του/της.
M-F	The accountant finished their work.	Ο/Η λογιστής/λογίστρια τελείωσε τη δουλειά του/της.

Table 20: Error instances produced by prompted GPT-4o for the ‘ambiguous / unambiguous [Non-binary]’ sentences.

I Gender Distribution of Google Translate, DeepL and GPT-4o in Unambiguous Cases

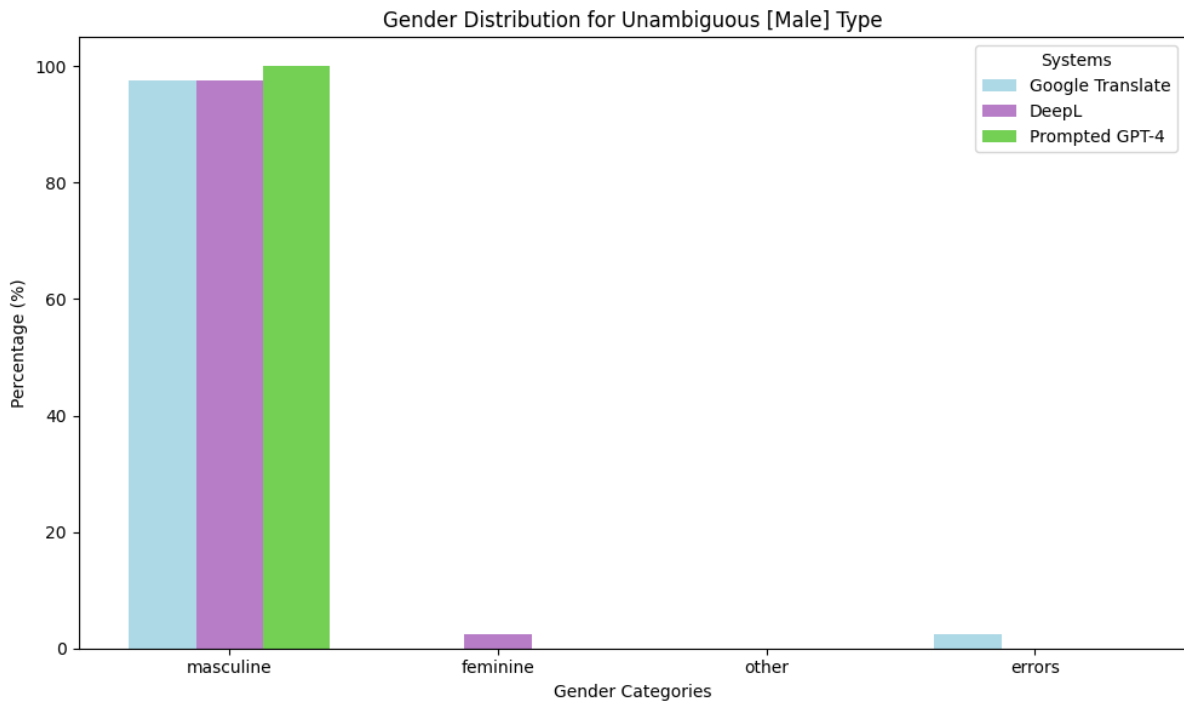


Figure 11: Gender distribution of translations for **masculine gender-unambiguous** sentences produced by Google Translate, DeepL, and (prompted) GPT-4o.

Gender Category	Google Translate	DeepL	Prompted GPT-4o
masculine	39	39	40
feminine	-	1	-
other	-	-	-
errors	1	-	-
Total	40	40	40

Table 21: Absolute counts for gender distribution of translations for **masculine gender-unambiguous** sentences produced by Google Translate, DeepL and prompted GPT-4o.

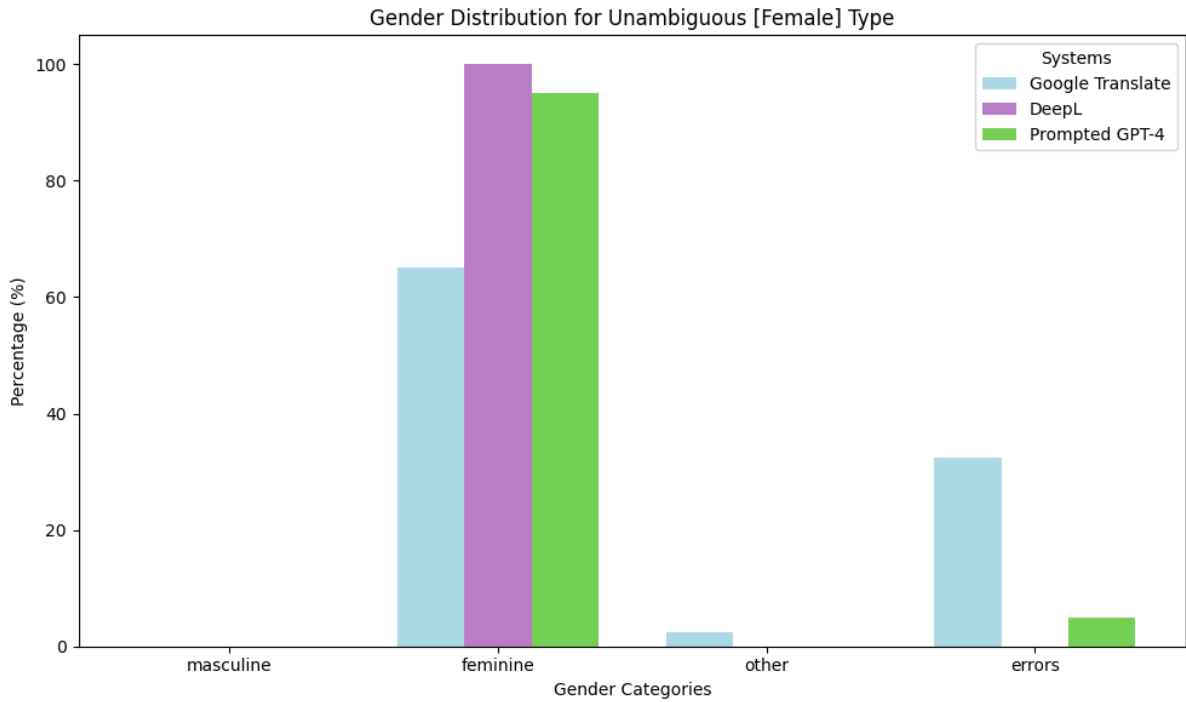


Figure 12: Gender distribution of translations for **feminine gender-unambiguous** sentences produced by Google Translate, DeepL, and (prompted) GPT-4o.

Gender Category	Google Translate	DeepL	Prompted GPT-4o
masculine	-	-	-
feminine	26	40	38
other	1	-	-
errors	13	-	2
Total	40	40	40

Table 22: Absolute counts for gender distribution of translations for **feminine gender-unambiguous** sentences produced by Google Translate, DeepL and prompted GPT-4o.