Faculty of Arts

Master’s Thesis

Digital Text Analysis

Gender Bias in English-to-Greek Machine Translation

Eleni Gkovedarou

Supervisor: Prof. Dr. Luna De Bruyne
Co-supervisor: Prof. Dr. Joke Daems

Assessor: Marika Fox

University of Antwerp

Academic year 2024–2025

The undersigned, Eleni Gkovedarou, student of the Master program in Digital Text Analysis at the University of Antwerp, declares that this thesis is completely original and exclusively written by herself. For all information and ideas derived from other sources, the undersigned has referred to the original sources, both explicitly and in detail.

Gender Bias in English-to-Greek Machine Translation

Eleni Gkovedarou*
University of Antwerp

In recent years, 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. Specifically, we address three aspects of gender bias: i) male bias (defaulting to masculine forms), ii) occupational stereotyping (assigning gender based on occupational stereotypes), and iii) errors in anti-stereotypical translations (e.g. more frequent mistranslations of "female doctors" versus "male doctors"). We also explore the potential of prompted GPT-4o as a gender rewriter capable of mitigating gender bias by providing gendered and gender-neutral alternatives. To achieve this, we introduce GendEL, a manually crafted dataset of 240 gender-ambiguous and unambiguous English sentences that feature stereotypical occupational nouns and adjectives. We find that gender bias is persistent 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 the feminine gender-unambiguous sentences, they are far from producing gender-inclusive translations when the gender remains unspecified. In contrast, prompted GPT-4o emerges as a promising solution for gender-inclusive translations as it offers gendered and gender-neutral alternatives for most ambiguous sentences, though some residual biases remain evident. Finally, we discuss the limitations of the models and dataset, and we open source our contributions to encourage further research in English-to-Greek machine translation.

1. INTRODUCTION

1.1 Gender Bias in Machine Translation

Machine translation (MT) technology has become an essential tool across various fields, facilitating not only daily tasks like communication, but also more complex functions such as knowledge sharing and collaboration across diverse linguistic communities. MT approaches have evolved significantly and rapidly, leading to current state-of-the-art neural machine translation (NMT). With the use of deep neural networks and artificial intelligence, NMT systems (such as transformer-based systems) have managed

* Lange Winkelstraat 40, 2000 Antwerpen

to overcome the earlier statistical machine translation (SMT) models in both fluency and accuracy, which have been the dominant technology for a long time (Mohamed et al. 2021). Despite NMT's great success, these systems are known to make systematic errors in areas like translating human gender, which leads to translations that are inaccurate, ungrammatical or biased (Stanovsky, Smith, and Zettlemoyer 2019; Currey et al. 2022). Concern has grown in recent years over the susceptibility of those systems to translate based on gender stereotypes and the perpetuation of such biases via them, as they can seriously harm users and society (Savoldi et al. 2021).

A machine translation model is considered biased "when it *systematically* and *unfairly* discriminates against certain individuals or groups in favour of others" (Savoldi et al. 2021). Since machine learning algorithms are written by humans and rely on data that was created, collected, cleaned and stored by humans, human errors and biases inevitably influence the algorithms and their outputs (Farkas and Németh 2022). 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, Hardmeier, and Way 2018; Kocmi, Limisiewicz, and Stanovsky 2020). These biases reflect the gender stereotypes that are present in our society. As Saunders, Sallis, and Byrne (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, Limisiewicz, and Stanovsky 2020).

The present 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). The cross-linguistic differences between these two languages can lead to ambiguities that are difficult to resolve, especially for sentence-level MT systems (Vanmassenhove, Emmery, and Shterionov 2021), making it more complex to accurately assign gender or maintain gender neutrality. Gender ambiguity in English can lead to forced gender assignment in Greek, frequently opting for the masculine form. This automatic bias not only misgenders individuals, but also fails to accommodate non-binary identities, which can negatively impact their mental health and ability to function in the world (Pino and Edmonds 2024). By focusing on the mitigation of gender bias in machine translation, we aim to advocate for fairer and more inclusive translation systems.

Despite growing attention to this area of study, research specifically focusing on the Greek language remains limited. While existing studies examine the behaviour of translation technologies on high-resource grammatical gender languages (e.g. Italian, French, German), we aim to investigate a medium-resource language like Greek to increase the language coverage in this field. Our study addresses this gap by creating **GendEL**,[1] the first dataset of gender-ambiguous and gender-unambiguous English sentences alongside human-generated Greek translations offering gender alternatives.[2] These English sentences are then translated using two commercial MT systems, Google Translate[3] and DeepL,[4] and analysed for gender bias through quantitative and qualitative evaluation

1 GendEL = "Gender" + "EL"
2 https://github.com/elenigkove/genderbias_EN-EL_MT
3 https://translate.google.com/
4 https://www.deepl.com/en/translator

metrics. Additionally, we present a translation-rewriting solution tailored to Greek, leveraging a prompted large language model (LLM), GPT-4o[5] (OpenAI, 2024), to offer both gendered and gender-neutral alternatives. Our goal is to support bias mitigation in English-to-Greek translations by exploring the ability of GPT-4o to provide not only accurate gender assignments, but also gender rewrites when needed.

1.2 Research Question and Hypotheses

In this study we investigate how commercial MT systems handle gender in English-to-Greek translation, a language pair that has been largely neglected so far. We explore the extent to which commercial MT systems, specifically Google Translate and DeepL, exhibit gender bias and whether a prompted LLM, such as the advanced GPT-4o, could be effective in fixing this issue. On the basis of prior literature and work conducted on other language pairs, we formulate the following two main hypotheses and sub-hypotheses:

1. **H1: Gender Bias.** Gender bias is expected to be present in the English-to-Greek translation outputs of Google Translate and DeepL. Specifically, we hypothesise the following patterns:
   (a) **H1a: Male Bias.** When translating gender-ambiguous English sentences, Google Translate and DeepL will frequently default to masculine-gendered forms, reflecting a male bias.
   (b) **H1b: Occupational Stereotyping.** MT systems are expected to reinforce stereotypical gender roles by translating male-biased professions (e.g. "doctor") with masculine forms and female-biased professions (e.g. "nurse") with feminine forms.

5 https://openai.com/index/hello-gpt-4o/

(c) **H1c: Anti-Stereotypical Gender Assignments.** More frequent errors or incorrect gender assignments are expected by the MT systems for sentences involving anti-stereotypical gender roles (e.g. "female doctors" or "male nurses") compared to stereotypical ones (e.g. "male doctors" or "female nurses").

2. **H2: GPT-4o on Bias Mitigation.** Using GPT-4o as a prompted LLM has the potential to mitigate gender bias by providing both gendered and gender-neutral options, thereby achieving more inclusive translations.

By testing the above, we aim to highlight the necessity for more inclusive and fairer translation systems, particularly for underrepresented language pairs like English-to-Greek.

1.3 Roadmap

This research is organised as follows: Section 2 provides an in-depth discussion structured in three parts: (a) the characteristics of Greek as a grammatical gender language, (b) gender-inclusive practices in Greek, and (c) a review of prior studies on gender bias in MT systems and the use of LLMs for bias mitigation. Section 3 outlines the methodology adopted for this research, detailing the dataset preparation, annotation, and translations performed by two commercial MT systems (Google Translate and DeepL) as well as the use of prompted GPT-4o. This section also explains the evaluation procedures employed for each (sub)hypothesis. Section 4 presents the results of the quantitative and qualitative analyses conducted for the (sub)hypotheses. Section 5 discusses the findings, offering interpretations of the models' performance. Section 6 acknowledges the limitations of the models and dataset, providing suggestions for future research and, finally, Section 7 concludes the study by summarising the key findings and highlighting its contribution to the ongoing efforts towards building fairer language technologies.

2. LITERATURE REVIEW

2.1 Greek as a Grammatical Gender Language

In Greek there are three declensions, signifying masculine, feminine, and neuter gender. Most linguistic components – including nouns, adjectives, passive participles, some pronouns and numerals – are gender marked. As a rich grammatical gender language, Greek typically uses masculine or feminine markings for human referents, while neuter is mostly assigned to inanimate objects. Nevertheless, there are certain grammatically neuter nouns, such as "το αγόρι" (the boy), "το άτομο" (the individual), or "το μέλος" (the member), that refer to humans. In general, nouns denoting male human beings are grammatically masculine and nouns denoting female human beings are grammatically feminine (Pavlidou, Alvanoudi, and Karafoti 2004). Alvanoudi (2015) points out that Greek's grammatical gender is restrictive, as gender marking must be encoded in most of the speakers' utterances. For example, the English sentence "This is a friend." has no overt expression of gender, however, in Greek the gender must be specified: "Αυτός/ή είναι ένας/μία φίλος/η" (This.MASC/FEM is a.MASC/FEM friend.MASC/FEM).

Given the scope of our study, the focus is primarily on the gender of occupational nouns. In such word types, gender marking is often overt, particularly in nominative singular forms, with gender signified by suffixes, e.g. "δάσκαλος" (teacher.MASC) and "δασκάλα" (teacher.FEM). However, gender marking can also be covert, especially in terms 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 (e.g. adjectives, pronouns, participles); for instance, "ο/η δικηγόρος" (the.MASC/FEM.SG lawyer.MASC.SG) 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 may be semantically linked to implicit connotations and is an indication of linguistic sexism in Greek (Hellenic Open University 2024). Several proposals (Triantafyllidis 1963; Tsopanakis 1982; Tsokalidou 1996; Gkasouka and Georgalidou 2014; Hellenic Open University 2024) have been made to feminise such 'common gender' terms in ways that align with the morphological and inflectional system of Greek while adhering to grammatical gender agreement, e.g. "η δικηγορίνα" (the.FEM.SG laywer.FEM.SG). However, despite a slow increase in acceptance and usage, most of these feminised terms have not become standardised in official language use.

Furthermore, only a limited number of professions has distinctly feminine suffixes, such as "μαία" (midwife), "καθαρίστρια" (cleaner), "δασκάλα" (teacher). 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 "δημαρχέσα" (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, Alvanoudi, and Karafoti (2004) examined all the nouns listed in the Dictionary of Standard Modern Greek and confirmed one of the fundamental claims of feminist linguists: the invisibility of women in language. Specifically, they found that when it comes to human reference, the masculine nouns are almost twice as many as the female ones, proving that the Greek vocabulary is male-dominated. This 'male-domination' is also exemplified in phenomena like the use of the generic masculine, where the masculine form of a term is used even when referring to mixed-gender groups, as is the case in most gendered languages (Savoldi et al. 2021). An example of this would be the sentence "Οι μαθητές είναι στο σχολείο" (the.MASC.PL students.MASC.PL are at school), which uses the masculine plural of "student" to include all genders. This practice leads to erasing women's and non-binary individuals' presence in language while representing male experience as "the default" (Mucchi-Faina 2005).

To this end, the Greek General Secretariat for Family Policy and Gender Equality (Gkasouka and Georgalidou 2014), and university entities (Petikas 2021; Hellenic Open University 2024) have released guidelines on gender-inclusive language that propose (1) the use of 'combined forms', where both feminine and masculine forms are used in full script ("ο καθηγητής / η καθηγήτρια" [the.MASC professor.MASC / the.FEM professor.FEM] or "η καθηγήτρια / ο καθηγητής"), (2) the use of 'combined suffixes', where the feminine or masculine form is abbreviated ("ο/η καθηγητής/τρια" or "η/ο καθηγήτρια/τής"), or (3) the use of exclusively feminine forms when referring to female entities. However, these strategies assume that gender is a dichotomous, binary social category, implying that references to the feminine and masculine are supposedly exhaustive (Ntouvlis 2020). This assumption excludes non-binary individuals, who may identify outside the traditional female and male categories. In English, for example, language reform has led to the adoption of the singular "they" and the creation of neopronouns (e.g. "ze", "zie", "xe", "ey" etc.), but Greek language reform has progressed at a slower pace, and equivalent options are not available, yet.

The most accessible gender-inclusive mechanism in Greek today, also included in the aforementioned guides, is gender-neutral language. This approach removes any gender markers altogether and thereby reduces potential discrimination (Karastergiou and Diamantopoulos 2024). Gender-neutral language can be achieved through various techniques, such as (a) passive syntax, (b) second-person plural, (c) imperatives, (d) circumlocution to avoid gender identification, and (e) neuter grammatical gender. Each technique has its limitations and is context-dependent, making application particularly challenging.

For our study on occupational nouns, a combination of (d) and (e) is the most viable neutralisation technique. As noted by Piergentili et al. (2023a), gender-neutral rephrasings and synonyms is a workable paradigm toward more inclusive MT when gender is unknown or simply irrelevant. For example, rather than using the gendered "καθηγητές/καθηγήτριες" (professors.MASC/FEM), we would employ multiword expressions like "το διδακτικό προσωπικό" (the teaching staff). To refer to an individual professor, a circumlocution that includes a neuter form such as "το μέλος του διδακτικού προσωπικού" (the.NEUT member.NEUT of the teaching staff) is preferred. 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 as self-identifiers, alongside neologisms like "το φίλο" (the.NEUT.SG friend.NEUT.SG), 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).

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 underrepresented 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.

2.3 Gender Bias in MT and 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, Avelar, and Lamb 2019; Cho et al. 2019; Stanovsky, Smith, and Zettlemoyer 2019; Kocmi, Limisiewicz, and Stanovsky 2020; Gonen and Webster 2020; Levy, Lazar, and Stanovsky 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, Hardmeier, and Way 2018). Additional methods include POS tagging (Elaraby et al. 2018), word-level gender tagging (Stafanovičs, Bergmanis, and Pinnis 2020; Saunders, Sallis, and Byrne 2020), fine-tuning (Saunders and Byrne 2020), or gender re-inflection of references into masculine/feminine forms such as Google Translate (Johnson 2020) and Fairslator[6] (Měchura 2022). As Savoldi et al. (2021) emphasise, there is no definitive, state-of-the-art solution for mitigating bias in machine translation; instead, these interventions typically address isolated aspects of the problem with targeted, modular solutions. It is worth noting that most of these studies largely operate within a binary framework, emphasising masculine and feminine forms into grammatical languages, which ultimately limits their inclusivity.

6 https://www.fairslator.com/

With the rapid advancement of large language models (LLMs), recent studies have examined their translation capabilities and potential for addressing gender bias. Ghosh and Caliskan (2023) investigated whether a prompted[7] GPT-3 perpetuates gender bias between English and Bengali, as well as five other low-resource languages (Farsi, Malay, Tagalog, Thai and Turkish) – all of which, except English, use gender-neutral pronouns. Translations from Bengali (or one of the five languages) into English showed that GPT-3 not only reinforces stereotypical gender assignments to certain gender roles (e.g. associating "doctor" and "going to work" with men, or "nurse" and "cooking" with women) but also systematically converts the gender-neutral pronouns of the source language into binary pronouns like "he" or "she" in English. Conversely, when translating from English into Bengali (or one of the five languages) the model completely fails to associate the English gender-neutral pronoun "they" into equivalent gender-neutral pronouns in the target languages, as it produces grammatically incorrect or nonsensical translations by treating "they" as a collective rather than an individual pronoun.

In a related study, Vanmassenhove (2024) explored how GPT-3.5 handles gender in English-Italian translation by simply prompting[8] it to translate sentences and by explicitly prompting[9] it to provide all the possible alternatives in terms of gender. Her findings reveal that in both prompting cases, the system indicates a strong male bias that becomes even stronger when explicitly asked to provide all possible gendered variations as, in numerous instances, the feminine gender or gender-neutral alternatives were entirely missing. It is important to also note that the corpus used in this study does

7 Prompt template (Bengali-English or English-Bengali): 'They are a + [insert occupational noun].'
8 Prompt: 'You are a Machine translation system. Can you translate the following sentence into Italian + [insert English sentence]'
9 Prompt: 'You are a Machine translation system. Can you translate the following sentence into Italian providing all the possible alternatives in terms of gender + [insert English sentence]'

not contain gender-neutral source sentences – those with the English gender-neutral pronoun "they".

In a further experiment, Lee et al. (2024) evaluated GPT-3.5 Turbo and Llama 2 70B Chat[10] for translations from English into Spanish, French and Italian, with the use of Gender-of-Entity (GoE) prompting. This approach instructs the model as a professional translator and allows the user to specify a preferred gender for each referent in the sentence,[11] producing controlled, gendered outputs. While this method showcases the potential of LLMs in handling gendered forms, it is currently limited to binary gender specifications.

Another study by Sánchez et al. (2024) tested the ability of Llama-7B[12] to provide two gender-specific translations for gender-ambiguous source sentences across twenty-five grammatical gender languages. By using in-context examples in the prompt (i.e. few-shot prompting),[13] they demonstrated a promising level of control over binary gender forms, raising the possibility that similar strategies might be applicable to non-binary alternatives as well. Building on this approach, Piergentili et al. (2024) followed the 3-shot format from Sánchez et al. (2024) to test LLMs for English-to-Italian translations, using the Neo-GATE dataset (built upon the GATE dataset by Rarrick et al. (2023)). Neo-GATE consists of English gender-ambiguous sentences with Italian references which only differ for the presence of either masculine, feminine or non-binary

10 https://huggingface.co/meta-llama/Llama-2-70b-chat-hf
11 User prompt: 'Translate the following sentence into [TGT LANG] ([GENDER ANNOTATION]): [SRC]'
12 https://huggingface.co/meta-llama/Llama-2-7b
13 Prompt: [EN] I have friends who are Hispanic people.
[ES] Tengo amigos que son personas hispanas.
[ES] Tengo amigas que son personas hispanas.

[EN] What do you think about ginger children?
[ES] ¿Qué piensas de los niños pelirrojos?
[ES] ¿Qué piensas de las niñas pelirrojas?

[EN] I have friends who are orphans.
[ES]
[ES]

words. Tailored specifically for Italian, the dataset incorporates non-binary structures, such as neomorphemes, special characters or symbols developed for inclusive language in Italian. Their findings indicated that few-shot prompting was more effective for controlling gender expression than zero-shot prompting, with GPT-4 and Mixtral[14] achieving the highest performance levels.

Within this landscape, LLMs have shown promise for offering a solution for gender inclusivity that goes beyond the binary framework of masculine and feminine. Savoldi et al. (2024) explored the neutral capabilities of GPT-4, by experimenting with three different prompts and few-shot exemplars on the English-Italian GeNTE dataset (Piergentili et al. 2023b). Their results revealed that, when prompted (few-shot prompting), GPT generated a notable amount of gender-neutral translations; however, in its baseline condition (zero-shot prompting), the model was found to be unsuitable for producing gender-neutral outputs. Although major MT providers have been criticised for persistent biased outputs in their systems, OpenAI claims to have implemented extensive bias mitigation measures. Yet, as these experiments suggest, the biases long observed and critiqued in commercial MT systems (e.g. Google Translate or MS Translator[15]) appear to persist even in advanced LLMs (Ghosh and Caliskan 2023).

Rarrick et al. (2024) developed a translation-rewriting solution with GPT-4, using chain-of-thought prompting (Wang et al. 2023), which involved explicitly providing the LLM with step-by-step reasoning and detailed clarifications in the examples. They examined translations from weakly-gendered languages (Turkish, Hungarian, Finnish and Persian) into English and the model was instructed to elicit three translation variants for each input sentence: one all-neutral, one all-female, and one all-male. This rel-

14 https://huggingface.co/docs/transformers/en/model_doc/mixtral
15 https://translator.microsoft.com/

atively simple rewriting task required adjusting pronouns to match the desired gender (he, she, they) and modifying verb forms (singular or plural), while leaving the rest of the sentence unchanged. The results indicate that while the model demonstrated high accuracy in rewriting pronouns, its performance declined when tasked with adjusting gendered nouns, showing a limitation in handling complex gender adjustments.

In conclusion, while significant progress has been made in developing strategies for mitigating gender bias in MT and LLMs, prior research has predominantly focused on binary frameworks, creating masculine and feminine alternatives without extending to gender-neutral expressions. This binary perspective limits the inclusivity of those systems, especially in grammatical gender languages, which could allow or evolve towards a spectrum of gender expressions. Greek, in particular, is a language that has remained notably underrepresented in these studies. To the best of our knowledge, this study is the first to investigate the ability of two commercial MT systems and a prompted LLM to accurately handle gender-ambiguous and gender-unambiguous sentences with a specific focus on stereotypically gendered occupational nouns and adjectives in this language pair. Additionally, we provide GendEL, the first English-Greek bilingual test set, with (human) annotations as well as gendered and neutral translations, intended as a gold standard for assessing these systems' performance and guiding future evaluation in this language pair.

3. METHODOLOGY

3.1 Dataset Preparation

For the purposes of our study, we constructed GendEL, a manually crafted dataset of 240 gender-ambiguous and gender-unambiguous English sentences. The dataset is based on a list of 40 occupational nouns. For each occupation, we created a subset of six sentences: one baseline sentence ('ambiguous base') and five variations. The baseline template follows the structure: *The [OCCUPATION] finished the work.*[16] Each variation modifies this template in a specific way, such as by adding a gender-biased adjective or a pronoun, so as to ensure a variety of gender contexts.

The occupational nouns were selected from the list collected by Troles and Schmid (2021) based on data from the US Bureau of Labor Statistics (2019) that provided gender distribution for various occupations. Accordingly, we classify an occupation as stereotypically male (male-biased), if the majority (> 50%) of workers in this field are men (e.g., 93% of carpenters are men), or as stereotypically female (female-biased), if most of the workers are women (e.g., 80% of librarians are women). Using this criterion, we constructed 20 subsets focusing on male-biased occupations, and 20 subsets focusing on female-biased occupations (Table 1).

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 1
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).

16 This template sentence was inspired by the dataset created by Saunders and Byrne (2020), which includes (binary) gendered examples such as "The actor finished her work" and "The actor finished his work". Their approach focused on leveraging a small dataset for transfer learning to improve gender debiasing in translation.

To enrich the scope of the study, we constructed five additional sentence types deriving from the 'ambiguous base', resulting in six types overall (examples are provided in Table 2):

- **Ambiguous + male-biased adj.**: A gender-ambiguous sentence containing a male-biased adjective.
- **Ambiguous + female-biased adj.**: A gender-ambiguous sentence containing a female-biased adjective.
- **Unambiguous [Male]**: Gender ambiguity is resolved with the use of a masculine pronoun.
- **Unambiguous [Female]**: Gender ambiguity is resolved with the use of a feminine pronoun.
- **Ambiguous / unambiguous [Non-binary]**: A sentence containing the singular pronoun "they", which makes the sentence either gender-ambiguous (the gender of the referent is purposefully omitted or undefined) or gender-unambiguous (singular "they" used to refer to a non-binary individual).

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

Table 2
Examples of sentence types representing a subset (for occupational noun "assistant") from GendEL.

The gender-biased adjectives were also sourced from Troles and Schmid (2021), who calculated the gender scores for several adjectives and gathered the ten most female-biased and ten most male-biased that could be combined naturally with occupational nouns (Table 3). Given that GendEL contains 40 subsets (one for each occupational noun) and there are 10 male-biased and 10 female-biased adjectives, each adjective needed to appear in four different subsets in order to ensure an even distribution. Within each subset, one male-biased and one female-biased adjective were randomly assigned to avoid systematic biases and ensure a diverse representation of gender-biased adjectives across the dataset.

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 3
Male-biased and female-biased adjectives included in GendEL.

Each of these sentence types were designed to explore how occupational nouns and certain contextual modifiers influence gender assignments during translations. To maintain the integrity of the analysis, we intentionally decided to keep the sentences short and simple, minimising linguistic diversity. This ensured that there were no factors influencing the referents' gender other than the investigated words, i.e. the occupational nouns, pronouns, and adjectives.

As part of the dataset preparation, these sentences were manually translated into Greek by the author of this study. Gender-ambiguous sentences were translated into three alternative translations (masculine, feminine, and neutral forms), whereas only one correct translation was produced for the gender-unambiguous sentences. For the 'ambiguous / unambiguous [Non-binary]' sentence type, we consider that there is only one correct translation which is a gender-neutral translation. In Greek, this is the standard approach when the source purposefully omits gender or refers to a non-binary individual, while any masculine or feminine representations would be considered non-inclusive.

It is important to note that while these human translations serve as a foundational baseline, they were not validated by professional translators or annotators. Therefore, they were not included in the evaluation process of our study but are intended as a resource for future research in order to establish a formal benchmark. A sample of GendEL, including subsets for a male-biased occupation and a female-biased occupation, is provided in Appendix A. Table 4 below summarises the key characteristics of GendEL:

Category	**Count**
Total sentences	240
Total subsets	40
Male-biased occupations	20
Female-biased occupations	20
Male-biased adjectives	10
Female-biased adjectives	10
Sentence types per subset	6
Gender-ambiguous sentences per subset	4*
Gender-unambiguous sentences per subset	3*

* The 'ambiguous / unambiguous [Non-binary]' sentence type is included in both gender-ambiguous and unambiguous categories due to its dual nature.

Table 4
Overview of the GendEL dataset composition.

3.2 Translation Systems

We test two widely used commercial MT models: (1) Google Translate and (2) DeepL. Both of these systems have implemented a feature that provides two outputs for short gender-ambiguous queries (Figure 1 and Figure 2).[17] However, while Google Translate

17 Screenshots taken on December 7, 2024.

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 (Figure 3 and Figure 4).[18]

Figure 1
Google Translate's output for the gender-ambiguous English sentence "I am a nurse" into Spanish. The system provides both feminine ("Soy enfermera") and masculine ("Soy enfermero") alternatives.

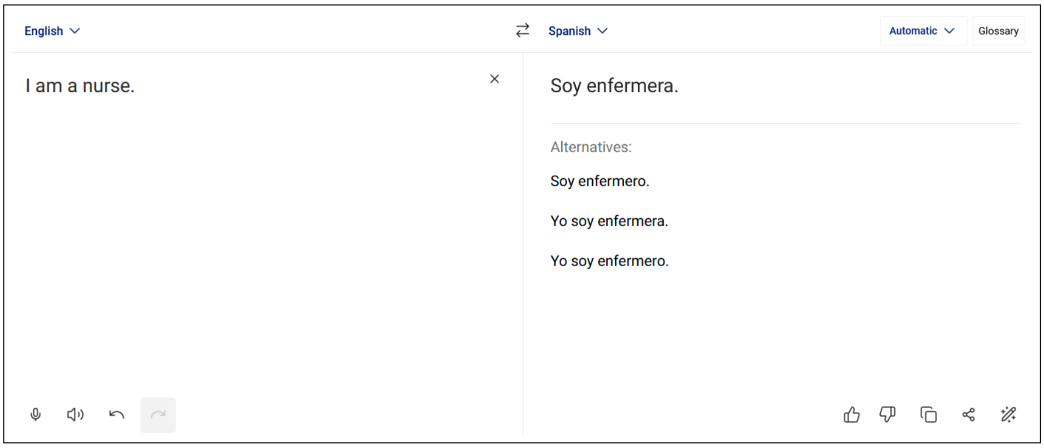

Figure 2
DeepL's output for the gender-ambiguous English sentence "I am a nurse" into Spanish. The main translation is the feminine form "Soy enfermera", while the alternatives include the masculine "Soy enfermero" and two variations with the pronoun "yo".

All sentences of GendEL were translated with both MT systems. For gender-ambiguous sentences we looked at the main output and all the different (gender-related)

18 Screenshots taken on December 16, 2024.

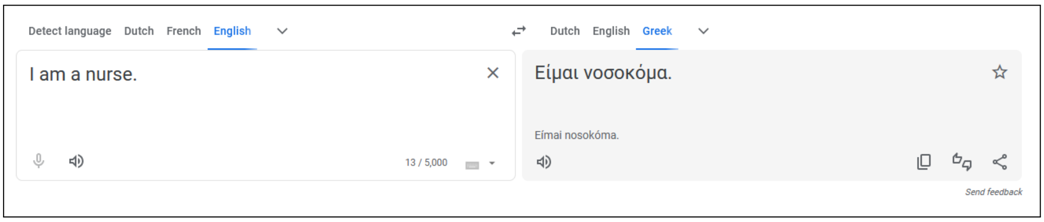

Figure 3
Google Translate's output for the gender-ambiguous English sentence "I am a nurse" into Greek. The translation is the feminine form "Είμαι νοσοκόμα" form without providing alternative gendered suggestions.

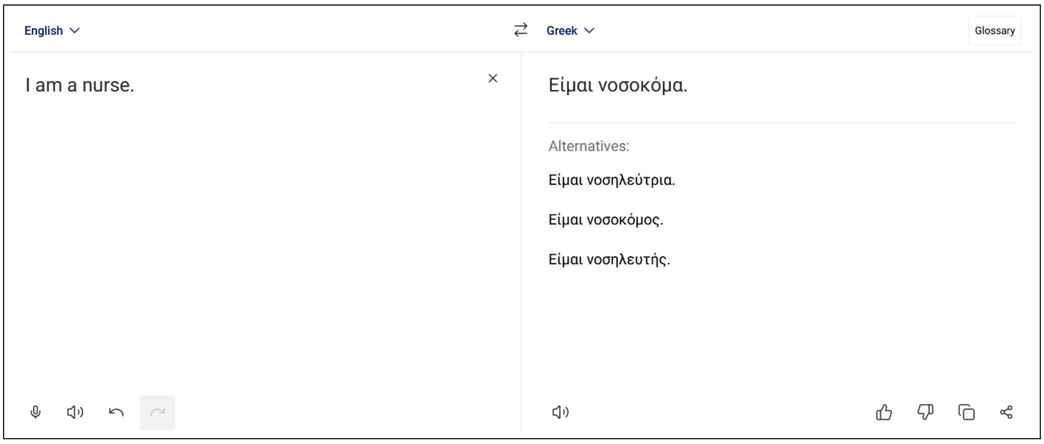

Figure 4
DeepL's translation for the gender-ambiguous English "I am a nurse" into Greek. While the main output uses the feminine form ("Είμαι νοσοκόμα"), alternative suggestions include both masculine ("Είμαι νοσοκόμος" and "Είμαι νοσηλευτής") and another feminine form ("Είμαι νοσηλεύτρια").

translations available from DeepL. However, in the case of gender-unambiguous sentences, where the assignment of gender is given, we evaluated only the main output produced by the system, as only one gendered output is needed in these cases. For instance, when translating "The nurse finished his work", if DeepL had a primary output with the masculine gender and an alternative associated with the feminine gender, only the main output was taken into account. This decision was taken in order to facilitate consistency and align with the nature of gender-unambiguous sentences, where a single gender assignment is expected. Although it would theoretically be possible to penalise alternative outputs for introducing superfluous gender forms, we opted to exclude this aspect in our evaluation to maintain a clear focus on the accuracy of the main output.

In addition to Google Translate and DeepL, we included GPT-4o into our analysis in order to examine its ability to handle gender ambiguity in translations from English into Greek. 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-4o[19] via the OpenAI API, integrating it with Python to automate the process of generating translations. Similarly to the approach of Rarrick et al. (2024), we aimed to encourage the model to produce gender rewrites, by using few-shot chain-of-thought prompting.

The LLM was provided with a detailed explanation of how to handle both gender-ambiguous and unambiguous source sentences (Appendix B). Specifically, for gender-unambiguous sentences, it was instructed to translate according to the source and provide only one translation, with an example included for clarity. In the case of gender ambiguity, it was explicitly asked to produce three variant translations: masculine, feminine, and neutral. Two examples were provided to illustrate this scenario; one with an adjective preceding the occupational noun and another without. Additionally, the prompt addressed the use of the singular pronoun "they" for non-binary individuals, instructing the LLM to generate a neutral Greek variant in such cases. Finally, it explicitly discouraged the use of techniques like combining masculine and feminine forms as a strategy for neutral translations (e.g. "ο/η επιθεωρητής/τρια" [the.MASC/FEM inspector.MASC/FEM]). By structuring the few-shot prompt in this way, we aimed to provide clear guidance to GPT-4o on the techniques for producing gender-inclusive translations in Greek for short queries like those included in GendEL.

19 Model version: gpt-4o-2024-08-06

For the finalisation of the dataset, we manually annotated the translations generated by the two MT systems and the 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)**: The translation is incorrect or nonsensical. Example: The laborer finished their work. = Οι εργάτες τελείωσαν τη δουλειά τους. (Backtranslation: The male laborers finished their work.) The singular form is not maintained in the output but is replaced with the plural.
- **error (2)**: The translation includes mixed genders. Example: The farmer finished her work. = Ο αγρότης τελείωσε τη δουλειά της. (Backtranslation: The male farmer finished her work.) The translation shows a mismatch where the masculine article and noun conflict with the feminine pronoun. While this output could potentially align with the source, it introduces a bias or inconsistency that makes the translation problematic.
- **error (3)**: This error signifies issues with the neutralisation techniques (incomplete or misleading). Example: The mechanic finished the work. = Το άτομο που εργάζεται ως μηχανικός τελείωσε τη δουλειά. (Backtranslation: The person who works as a mechanic finished the work.) The epicene noun "μηχανικός" (mechanic) implies a binary gender classification and does not reflect true gender neutrality. Thus, it is not a sufficient neutralisation strategy.

- **error (4)**: The adjective is missing from the output. Example: The perky lawyer finished the work. = Το άτομο που ασκεί τη δικηγορία τελείωσε τη δουλειά. (Backtranslation: The person who practices law finished the work.) "Perky" is omitted from the translation.

It is important to note that the annotation labels are mutually exclusive, meaning each translation is assigned only one label. The error label complicates gender classification, especially in 'error [1]', 'error [2]', and 'error [3]' cases. Although incorrect gender assignments are technically errors, they are not classified as such; instead, they are assigned the corresponding gender label. For example, if a gender-ambiguous "professor" or a gender-unambiguous "male professor" is translated as a "female professor", the output would be classified as 'F' rather than 'error [x]'. This decision is based on the aim of assessing the overall patterns of gender bias in the translations. Therefore, incorrect gender assignments are not categorised as errors, however, depending on the sentence type, we can identify which gender label is incorrect, enabling a clearer understanding of the bias.

3.3 Evaluation

The evaluation of the MT and LLM models was performed using a mixed-methods approach. Automatic evaluation metrics such as BLEU (Papineni et al. 2002) or TER (Snover et al. 2006) were deliberately not included in the study, due to their 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). They do this by comparing n-grams of the machine-translated text to those in a reference translation. Thus, whenever there is a deviation from the reference (e.g. stylistic differences, mistranslations, gender-related issues, etc.), they are all treated the same, regardless of their nature or significance. In other words, they report that there is a difference between the reference and machine-generated translation, but do not provide information about the type of difference. This would be especially problematic in the language pair we are examining due to the challenging properties of the Greek language, mainly with regards to its gender-neutral strategies. Therefore, we rely on human interpretation to assess the translations' validity and their alignment with gender-inclusive practices in Greek, ensuring a more accurate evaluation of the systems' performance.

3.3.1 Gender Bias Hypothesis (H1). To investigate the presence of gender bias in the MT systems, we examined three key patterns of bias using a mixed-methods approach that combined quantitative and qualitative metrics: male bias (H1a), occupational stereotyping (H1b), and errors in anti-stereotypical gender assignments (H1c).

Male Bias (H1a): 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[20] in the dataset, 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 (H1b): To explore occupational stereotyping, we assessed whether the MT systems reinforced traditional gender roles associated with specific professions (e.g. "male doctor", "female nurse"). Here, our focus is again exclusively on the gender-ambiguous sentences. We calculated the MT systems' frequency of stereotyping based on the stereotypical gender of the occupational noun and examined the gender

20 'ambiguous base', 'ambiguous + male-biased adj.', 'ambiguous + female-biased adj.', 'ambiguous + unambiguous [Non-binary]'

distribution for male-biased and female-biased occupations separately. Furthermore, we performed statistical significance testing using Fischer's exact test. This test is particularly suitable for analysing categorical data, especially when some we are dealing with small counts. It measures whether an observed association between variables – in our case, gendered translations and the stereotypical bias of professions – is statistically significant rather due to random chance.

Anti-Stereotypical Gender Assignments (H1c): For H1c, we analysed the outputs of gender-unambiguous English sentences[21] and focused on the translation errors and incorrect gender assignments. Particularly, we compared the error rates between anti-stereotypical and stereotypical professions. For example, we investigated whether it was more likely for Google Translate or DeepL to generate erroneous translations when dealing with anti-stereotypical cases (e.g. "female doctor" or "male nurse"), compared to stereotypical ones (e.g. "male doctor", "female nurse"). Error rates were calculated and compared for both categories, followed by a Fischer's exact test for each case. In addition to the quantitative analysis, we conducted a qualitative review of errors in the single case where a statistically significant difference was observed, offering a deeper understanding of the nature of these errors.

3.3.2 GPT-4 on Gender Bias Mitigation Hypothesis (H2). The final part of our study included the evaluation of the prompted GPT-4o as a translation tool specifically customised to produce gender-inclusive outputs. To this aim, we calculated the gender distribution of gender across all the different sentence types of GendEL. This already allowed us to draw some preliminary conclusions regarding the effectiveness of the

21 'unambiguous [Male]', 'unambiguous [Female]'

model. For this phase, all error categories were merged into a single "error" label to reduce any potential noise in the analysis.

In the second phase, the focus was drawn upon the errors produced by GPT-4o. Specifically, we calculated the error distribution across the sentence types, providing an initial overview of the nature of these errors. To gain a better understanding, we performed a qualitative analysis which helped uncover potential factors that influence the performance of the model and highlighted areas where it deviated in terms of gender-inclusive practices. We will present the detailed results of these analyses in the following section, including the specific patterns observed and enabling a more in-depth discussion of the models' global performance.

4. RESULTS

A preliminary overview of the distribution of gender labels and error types across the three investigated models, reveals substantial differences. From the data in Table 5, it is apparent that Google Translate and DeepL exhibited a strong tendency to favour masculine forms across the whole dataset, with 158 (65.8%) and 152 (63.3%) masculine translations, respectively. The feminine translations for those systems were significantly fewer, with only 39 (16.2%) female gender assignments for Google Translate and 62 (25.8%) for DeepL.

On the contrary, prompted GPT-4o showed a more balanced approach, generating 40 (16.7%) masculine and 38 (15.8%) feminine translations, while also offering a notable number of 31 (12.9%) gender-neutral outputs. For 103 (42.9%) sentences it generated all three gendered alternatives ('M-F-N'). These results stand out against the gold standard statistics, which represent the idealised distribution of gendered and neutral translations, with equal proportions of masculine and feminine forms (16.7% each), and a balanced representation of neutral (16.7%) and alternatives (50%).

In terms of translation errors, Google Translate was the model with the highest numbers of errors: 32 (13.3%) translations included mixed genders, 9 (3.7%) were incorrect or nonsensical translations, and 1 (0.4%) translation omitted the adjective. The number of errors by DeepL was significantly lower, with only 3 (1.3%) mixed-gender and another 3 (1.3%) incorrect translations. Prompted GPT-4o also showed low rates of mixed-gender and mistranslation errors – 4 (1.7%) and 3 (1.3%), respectively – but it omitted gender-biased adjectives in 10 (4.2%) translations. 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 8 (3.3%) errors related to neutralisation techniques.

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%)	-

Table 5
Distribution of gender and error labels across the three systems and gold standards, with raw counts and proportions.

This surface-level analysis suggests that the MT systems are more likely to display gender bias, whereas prompted GPT-4o seems to have the potential to produce gender-inclusive translations. The gold standard column in Table 5 captures the discrepancy between observed and ideal distributions, particularly regarding the lack of gender-neutral forms and gendered alternatives in Google Translate and DeepL. The following subsections analyse particular trends of gender bias and errors to evaluate the validity of the (sub)hypotheses outlined in Section 1.2.

4.1 Hypothesis 1 – Gender Bias: Findings

4.1.1 Male Bias (H1a). As outlined earlier, the first pattern of gender bias for exploration was the presence of male bias in the outputs of Google Translate and DeepL. An analysis of their translations for the 160 gender-ambiguous English sentences revealed a clear tendency towards male bias. Figure 5 illustrates the gender distribution for the two MT systems when translating source sentences that do not indicate any gender.[22] With masculine forms making up 74.4% of Google Translate's output and 70.6% of DeepL's, it is evident that these systems tend to favour masculine over feminine or neutral forms, aligning with common findings about male default bias in MT systems.

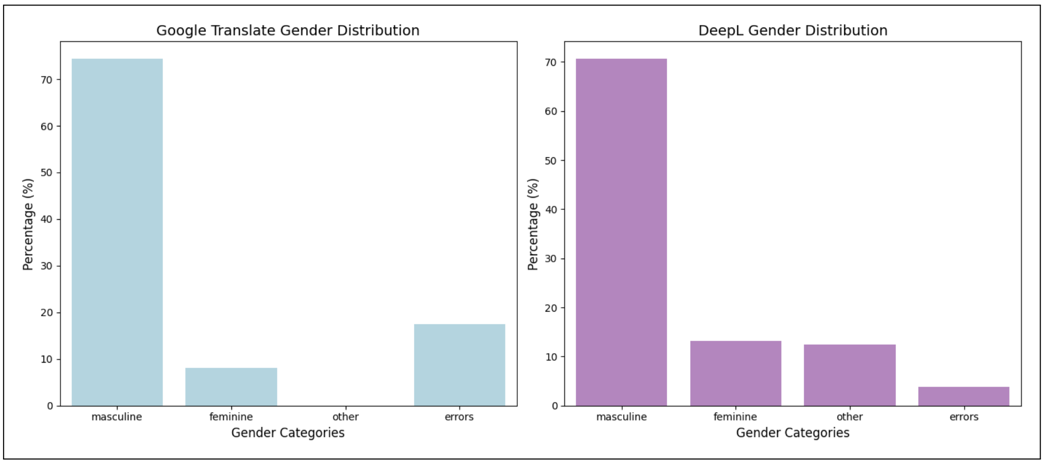

Figure 5
Gender distribution of translations for **gender-ambiguous** sentences produced by Google Translate and DeepL. Absolute counts are provided in Appendix C.

22 Outputs 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 in the illustrated distribution bar plots.

Among non-masculine outputs, the findings show an overall low percentage of feminine forms for both systems (13.1% for DeepL and 8.1% for Google Translate), while errors are significantly higher for Google Translate – 17.5% against DeepL's 3.7%.

A closer look at the errors, revealed that most of the errors by Google Translate stemmed from 'error [2]' (i.e. mixed-gender representations), which accounted for 13.1% of the total outputs. Notably, all of these 'error [2]' instances appeared in the 'ambiguous / unambiguous [Non-binary]' category, suggesting that the model particularly struggled in handling the gender-neutrality or non-binarity expressed by the singular pronoun "they". Specifically, these errors indicate that the model treated "they" as a collective pronoun and defaulted to the masculine form when it came to the profession, and the collective "they" when translating "their work", failing to correlate the gender-neutral pronoun with gender-neutral solutions in Greek.

On the contrary, DeepL produced far fewer errors overall, with 1.8% of outputs classified as mistranslations ('error [1]') and another 1.8% as mixed-gender outputs ('error [2]'). Interestingly, despite the low rate of errors for this system, the errors were again found in the 'ambiguous / unambiguous [Non-binary]' sentence type, suggesting that DeepL also struggles with sentences involving gender-neutral or non-binary pronouns.

Regarding gender-inclusive outputs, neither system performed well. Google Translate did not produce any gender-neutral or inclusive translations. DeepL performed slightly better, providing alternate translations for 12.5% of the gender-ambiguous sentences. For the majority of these cases, this involved generating one masculine and one feminine translation, while for only one gender-ambiguous sentence, DeepL produced a masculine translation paired with a gender-neutral alternative. These results reveal slightly, though not significantly, better performance of DeepL over Google Translate in producing gender alternatives, however, the translations remained within a binary construct and were not fully gender-inclusive. Overall, both MT systems exhibit a notable male bias in their handling of gender-ambiguous English sentences, thus the first subhypothesis is **accepted**.

4.1.2 Occupational Stereotyping (H1b). The second investigated pattern of gender bias was occupational stereotyping in gender-ambiguous source sentences. Figure 6 and Figure 7 present the gender distributions for Google Translate and DeepL when the occupational noun was stereotypically male and female, respectively.

With masculine outputs reaching 82.5% (Google Translate) and 86.2% (DeepL) of the translations when the occupation is male-biased, and 66.2% (Google Translate) and 55% (DeepL) when the occupation is female-biased, the results show that the masculine forms dominate the translations for both systems, regardless of the stereotypical gender of the occupational noun in the sentence. This observation aligns with the findings of the previous section regarding the persistent male bias in MT systems.

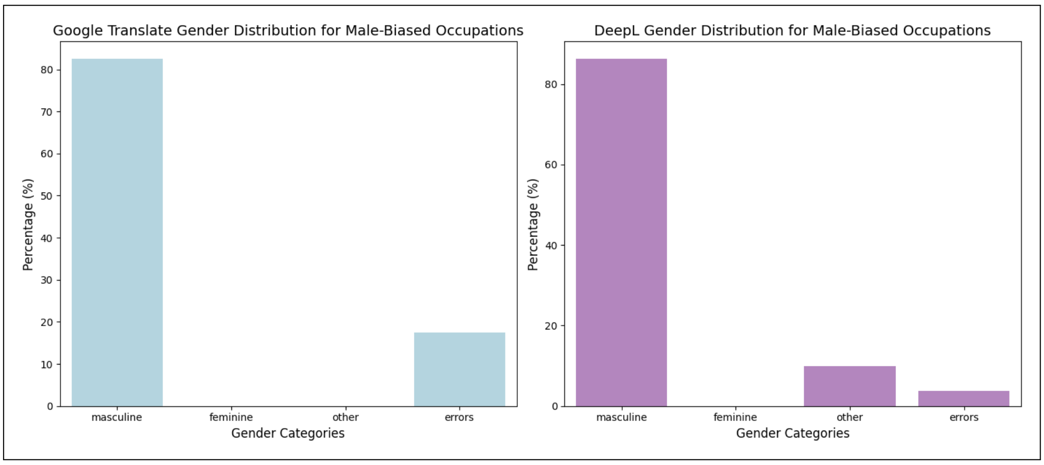

Figure 6
Gender distribution of translations for **stereotypically male occupations** in **gender-ambiguous** sentences, produced by Google Translate and DeepL. Absolute counts are provided in Appendix C.

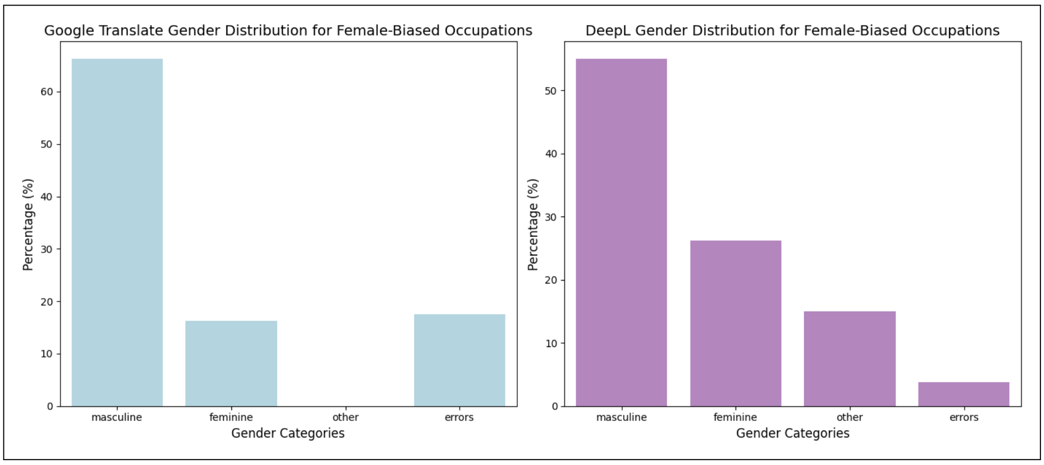

Figure 7
Gender distribution of translations for **stereotypically female occupations** in **gender-ambiguous** sentences, produced by Google Translate and DeepL. Absolute counts are provided in Appendix C.

A closer examination of the results reveals an interesting pattern in the feminine gender category. 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 indicates that while masculine remains the default, MT systems are influenced by societal stereotypes, associating feminine forms more frequently with traditionally female-biased professions.

To further investigate this phenomenon, a Fischer's exact test was conducted to evaluate whether there is a statistically significant association between occupational stereotypes and gender outputs. The test results for both systems (Appendix D) confirmed that the stereotype of the occupation significantly impacts the translation gender, with feminine forms significantly more likely to appear for stereotypically female occupations than stereotypically male ones. Therefore, H1b is **partially accepted**, as the systems showed a strong association between occupational stereotypes and the gender of the translations, but the masculine form remained the overall default.

4.1.3 Anti-Stereotypical Gender Assignments (H1c). The third and final pattern of gender bias investigated in this study relates to the accuracy of gender assignments when MT systems translate sentences involving anti-stereotypical gender roles compared to stereotypical ones. In simpler terms, the systems were expected to generate more frequent errors or incorrect gender assignments when translating "female doctors" or "male nurses" than "male doctors" or "female nurses". For this type of bias, only gender-unambiguous sentences were examined, as outlined in Section 3.

Figure 8 and Figure 9 below illustrate the gender distribution of Google Translate and DeepL for stereotypical gender assignments. As expected, for male-biased occupational nouns, both systems successfully translated the 'unambiguous [Male]' sentences, where the gender ambiguity was resolved with the use of a masculine pronoun ("his work") (Figure 8). The majority of translations correctly assigned masculine gender, with only one mistranslation from Google Translate: the word "mover" was translated as "μετακινούμενος" instead of "μεταφορέας". This mistranslation reflects a lexical error rather than a gender-related issue, as "μετακινούμενος" refers to someone being "moved" rather than the profession of a "mover".

For female-biased occupational nouns found in 'unambiguous [Female]' sentences ("her work") (Figure 9), DeepL demonstrated consistently accurate performance, correctly assigning feminine gender to all translations. Google Translate, however, produced three erroneous outputs: two with mixed-gender representation and one with a lexical error in the occupational noun. Specifically, the system translated "female baker" as "φούρναρη", which is a non-existent (feminised) form of the noun. This potentially

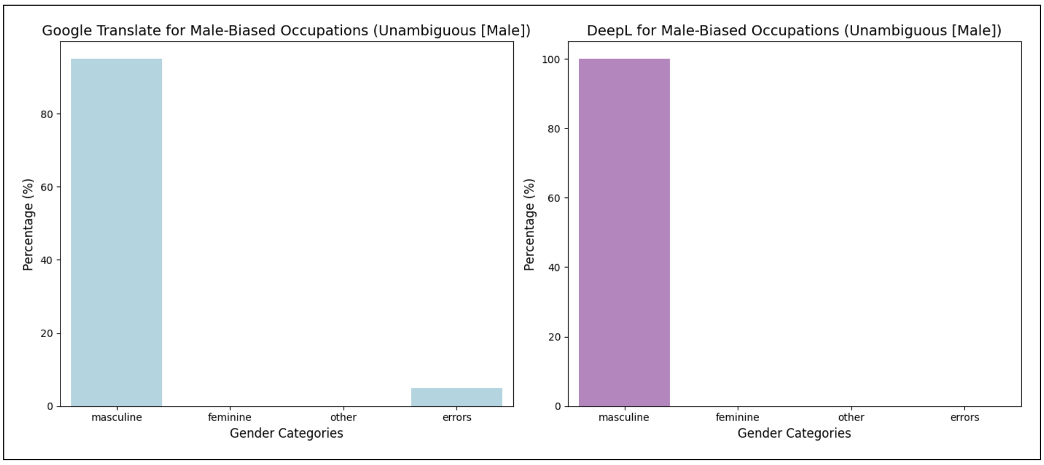

Figure 8
Gender distribution of translations for **stereotypically male occupations** in **masculine gender-unambiguous** sentences (stereotypical case), produced by Google Translate and DeepL. Absolute counts are provided in Appendix C.

indicates not only a grammatical error but also the system's difficulty in handling feminine forms for some professions.

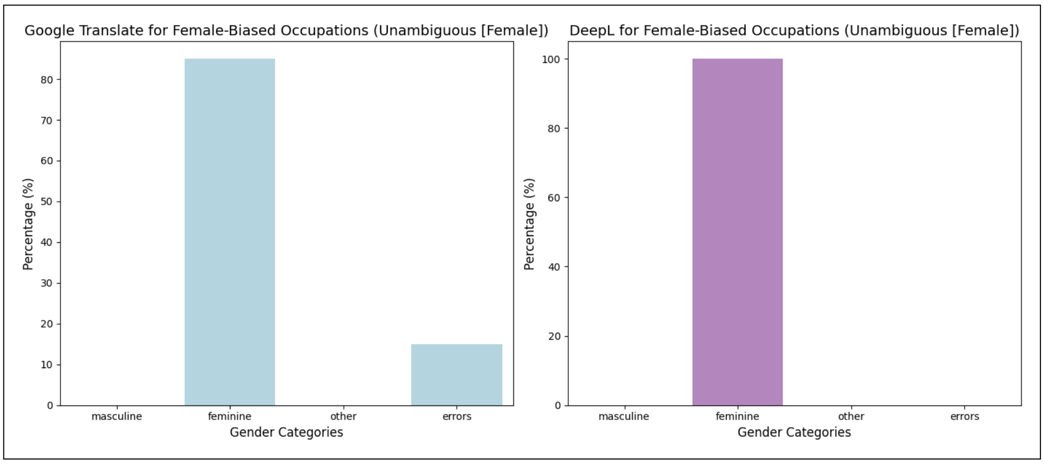

Figure 9
Gender distribution of translations for **stereotypically female occupations** in **feminine gender-unambiguous** sentences (stereotypical case), produced by Google Translate and DeepL. Absolute counts are provided in Appendix C.

According to our subhypothesis, we expected more incorrect gender assignments or errors in anti-stereotypical sentences. Nevertheless, Figure 10 does not seem to support this claim, since for ’unambiguous [Male]’ sentences containing female-biased occupations, Google Translate produced accurate masculine translations in all cases. DeepL also performed well, with only one incorrect gender assignment. In this instance, the occupation "housekeeper" was translated into the feminine form, despite the explicit masculine pronoun in the English source.

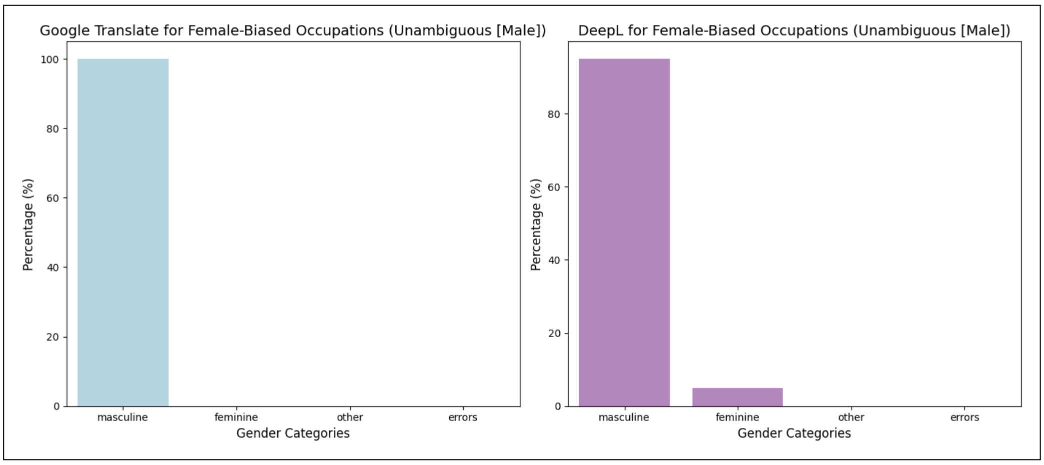

Figure 10
Gender distribution of translations for **stereotypically female occupations** in **masculine gender-unambiguous** sentences (anti-stereotypical case), produced by Google Translate and DeepL. Absolute counts are provided in Appendix C.

In contrast, for ’unambiguous [Female]’ sentences containing male-biased professions (Figure 11), 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, and one instance was labelled as ’other’.

A closer qualitative analysis of these outputs showed that the ’other’ case involved a gender-neutral translation. In particular, "female guard" was translated as "η φρουρά" (backtranslation: "the guard", as in a neutral term used to refer to the role without specifying gender), which is neutral in Greek and does not indicate gender. As for the significant number of errors, the majority were of the mixed-gender type, such as "The

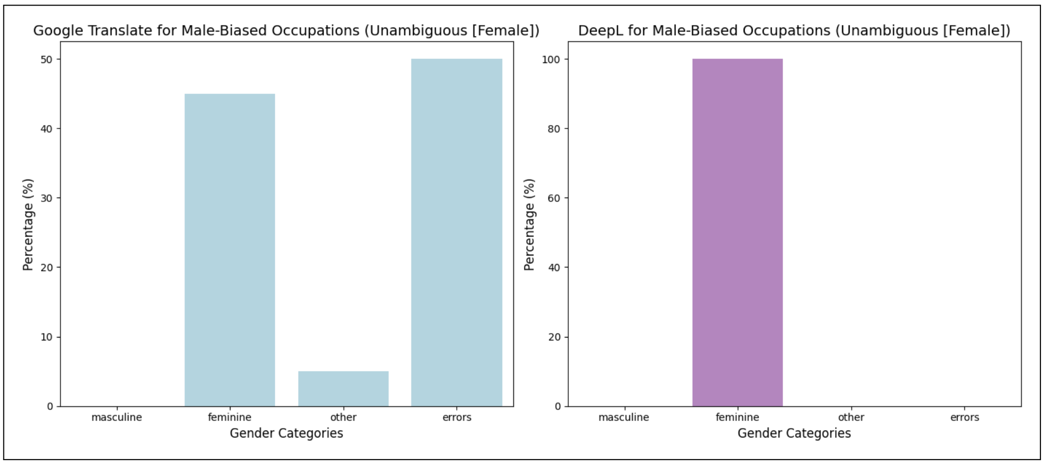

Figure 11
Gender distribution of translations for **stereotypically male occupations** in **feminine gender-unambiguous** sentences (anti-stereotypical case), produced by Google Translate and DeepL. Absolute counts are provided in Appendix C.

male driver finished her work", where conflicting gender markers created incoherence. Additionally, one instance involved a severe mistranslation (lexical error): "construction worker" was translated as "η οικοδομή" (the building). The grammatical gender of this noun is feminine, suggesting the model, in an attempt to assign feminine form to the output, used an incorrect term.

Out of all investigated patterns, the only statistically significant difference occurred between the stereotypical and anti-stereotypical groups in Google Translate, specifically for the ’unambiguous [Female]’ sentences. The results of the Fischer’s exact testing revealed that Google Translate is more likely to produce feminine forms when the occupation is stereotypically female (stereotypical case, e.g. "female nurse") compared to when it is stereotypically male (anti-stereotypical case, e.g. "female doctor"). This also suggests that Google Translate struggles particularly with translating male-biased professions into feminine forms, even when an explicit feminine pronoun is present in the source text.

Overall, DeepL outperformed Google Translate across feminine- as well as masculine-gendered English sentences. Almost regardless of the context being stereotypical or anti-stereotypical, in the vast majority of cases the system assigned the correct gender. Hence, H1c is **partially accepted**, revealing some issues specific to Google Translate's processing of anti-stereotypical gender assignments, particularly in male-biased contexts.

4.2 Hypothesis 2 – GPT-4o on Gender Bias Mitigation: Findings

4.2.1 Quantitative Analysis. The performance of the prompted GPT-4o showed promising results overall. An overview of the gender distribution across the different sentence types is presented in Figure 12. For the **'unambiguous [Male]'** sentences, the model achieved a 100% success rate, correctly translating all instances into the masculine form, suggesting that it can handle masculine forms with high precision when the gender is explicitly specified. For the **'unambiguous [Female]'** sentences, the model successfully generated feminine forms for 95% of the cases while the remaining 5% in this category involved errors. A detailed analysis of these error types will follow in Section 4.2.2.

Furthermore, for the **'ambiguous base'**, **'ambiguous + male-biased adj.'**, and **'ambiguous + female-biased adj.'** sentences, the model showed high accuracy in detecting gender ambiguity in the source text. It generated three correct gender alternatives ('M-F-N') as instructed in the prompt, with success rates reaching up to 92.5%, 80% and 85%, respectively. The remaining cases were classified as errors.

Finally, for the **'ambiguous / unambiguous [Non-binary]'** sentences, the model successfully produced gender-neutral translations using a circumlocution involving the neuter form "το άτομο που" (the person who). However, 7.5% of translations in this category were labelled as 'M-F', referring to outputs that included double forms (e.g. "ο/η λογιστής/λογίστρια" – the.MASC/the.FEM accountant.MASC/accountant.FEM),

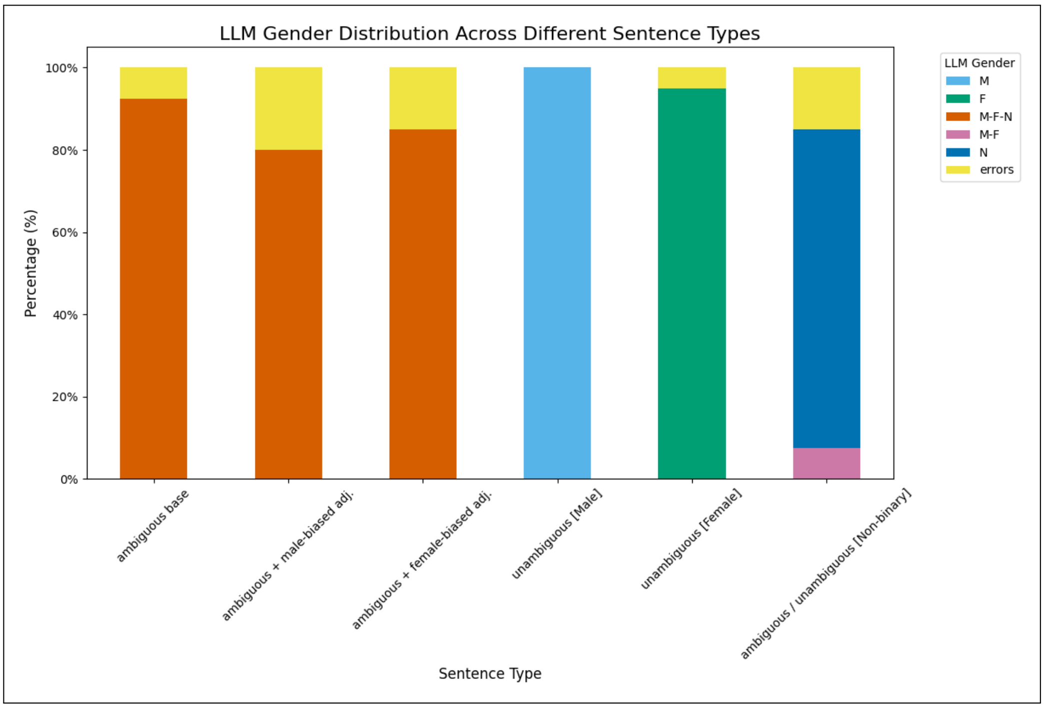

Figure 12
GPT-4o gender distribution across all sentence types. Absolute counts and proportions are provided in Appendix E.

despite being explicitly instructed to avoid them, with an example provided in the prompt. Additionally, 15% of the translations in this category were classified as errors.

The above results show that the 'unambiguous [Male]' sentence type obtained the highest accuracy, followed by the 'unambiguous [Female]' type. 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 the prompted GPT-4o with Google Translate and DeepL on gender-unambiguous sentences, revealed that the LLM outperformed the other two MT systems with 100% success rate in masculine gender-unambiguous sentences (Figure 13). Interestingly, in feminine unambiguous sentences

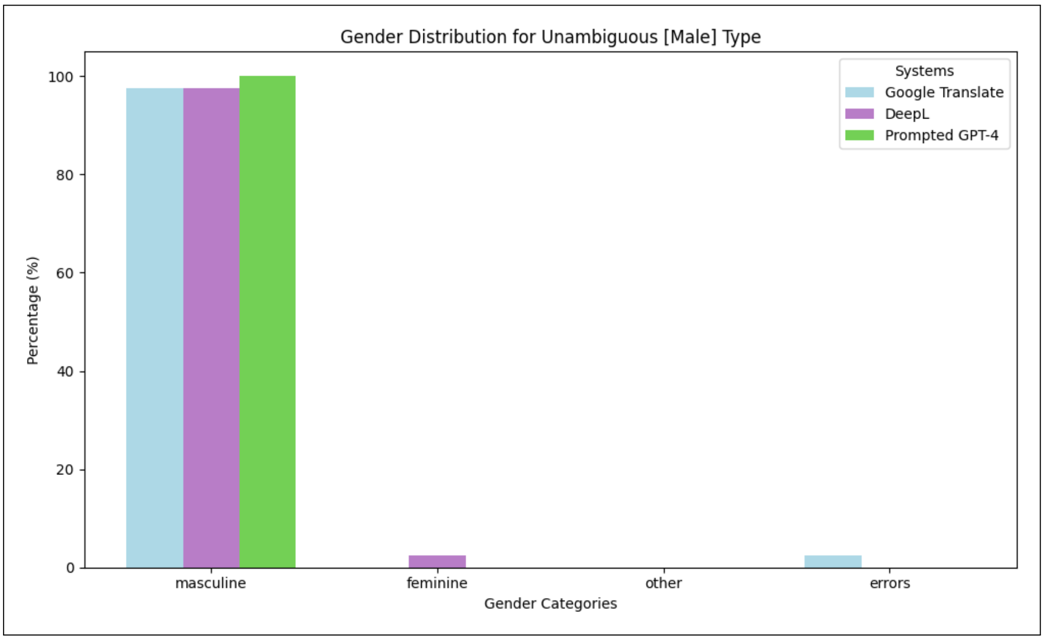

Figure 13
Gender distribution of translations for **masculine gender-unambiguous** sentences produced by Google Translate, DeepL, and (prompted) GPT-4o. Absolute counts are provided in Appendix C.

(Figure 14), 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 challenges in fully overcoming biases ingrained in the model's training data, even when customised to address such issues.

Overall, the performance for gender-ambiguous sentences demonstrates the model's ability to handle gender ambiguity effectively, generating three 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 (combining 'M-F'

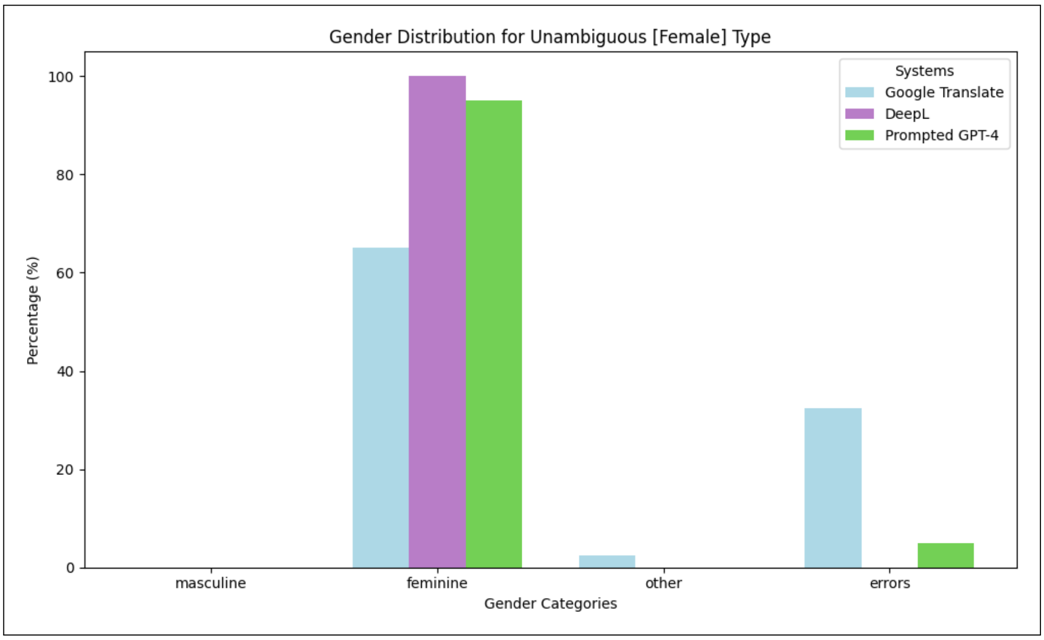

Figure 14
Gender distribution of translations for **feminine gender-unambiguous** sentences produced by Google Translate, DeepL, and (prompted) GPT-4o. Absolute counts are provided in Appendix C.

and 'errors' labels), which may reflect the limitations of the model in producing accurate gender-neutral language.

4.2.2 Qualitative Analysis of Errors. A closer investigation of the errors for each sentence category shows interesting information regarding the inaccuracies produced by the prompted GPT-4o. As a quick recap, 'error [1]' are outputs that are incorrect or nonsensical, 'error [2]' involves mixed-gender forms, 'error [3]' captures issues with neutralisation, and 'error [4]' were cases where an adjective was missing in the translations. Incorrect gender assignments will be reported with their respective gender label. While proportions were used for reporting the overall results of the investigated systems, we now focus on raw error counts to facilitate clearer interpretation, especially given the low frequency of errors.

Figure 15 illustrates the distribution across all sentence types, highlighting how specific error categories vary depending on the sentence type. A preliminary overview reveals that the 'ambiguous / unambiguous [Non-binary]' category spans three error types, suggesting that the model faced significant challenges in translating these sentences. In the following paragraphs, we qualitatively examine each sentence type and the associated error categories aiming to uncover patterns and potential causes.

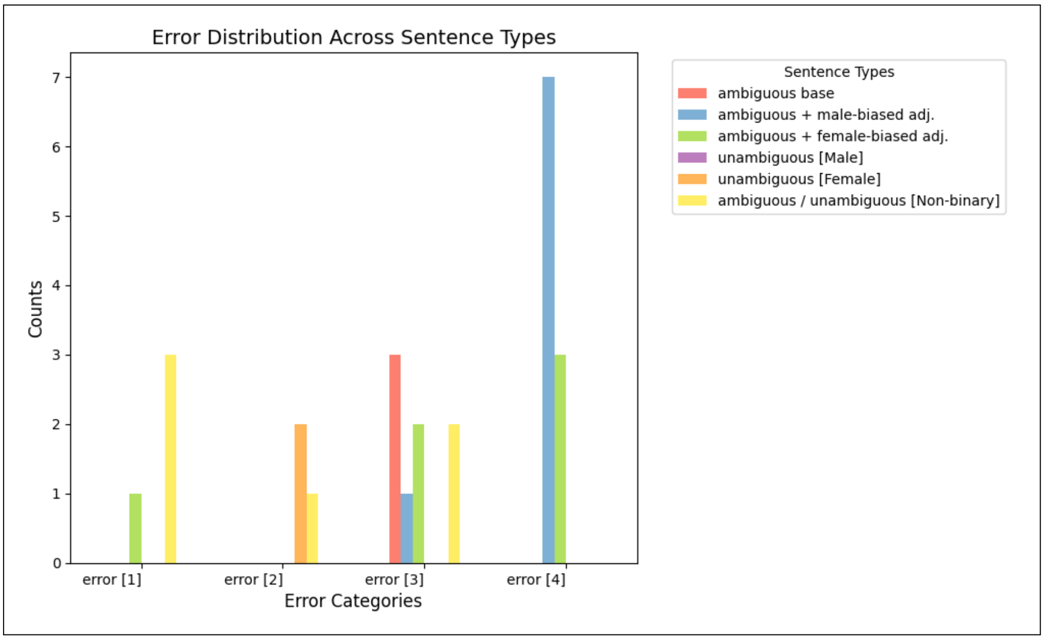

Figure 15
Error distribution of prompted GPT-4o across all sentence types.

Ambiguous base

For the 'ambiguous base' type, three instances were labelled as 'error [3]', indicating challenges in producing fully gender-neutral language. Specifically, the model consistently produced three gender alternatives for each source sentence, in which the masculine and feminine versions were accurate. However, the neutral form presented issues, as seen in Appendix F. The problematic outputs attempted to use the neuter circumlocution "το άτομο που" (the person who) but were paired with epicene nouns, such as "μηχανικός" (mechanic), "σερίφης" (sheriff), and "διευθύνων σύμβουλος" (CEO). As discussed in Section 2, while epicene nouns are used for both male and female genders, they rely on a masculine suffix and imply a binary construct, thereby excluding non-binary individuals. Thus, they are not considered a satisfactory neutralisation technique. This limitation highlights the model's struggle to move beyond traditional gender constructs and generate more inclusive translations.

Ambiguous + male-biased adj.

Regarding the 'ambiguous + male-biased adj.' category, seven outputs were labelled as 'error [4]' and one as 'error [3]'. It is worth noting that, once again, all the errors occurred in the gender-neutral alternatives provided by the model, alongside the correct male and female versions. The examples in Appendix F demonstrate the types of errors observed for this category. Similar to the 'ambiguous base' sentences, there was an instance where the model attempted to neutralise gender by using the neuter circumlocution "το άτομο που" (the person who), followed by the epicene noun "μηχανικός" (mechanic). Most issues in this category, however, 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 assumed 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.

'Error [4]' also appears in the 'ambiguous + female-biased adj.' category, accounting for three translations, while two outputs were labelled as 'error [3]' and only one as 'error [1]'. Once again, the model shows difficulty in correctly neutralising sentences without relying on an epicene noun, and often omits descriptive details from the source text (i.e. female-biased adjectives). Interestingly, one instance was classified as incorrect translation due to an error involving the repetition of the word "που" ([the person] who). Notably, all errors occurred in the neuter alternative produced by the model, highlighting again its struggle with gender neutralisation. The errors for this sentence type are illustrated in Appendix F.

Unambiguous [Female]

For the 'unambiguous [Female]' type, errors were minimal, with two instance classified as 'error [2]' (Appendix F). This error shows a mismatch between the grammatical gender of the subject "sheriff" and "farmer" and the personal pronoun "her". 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]

Finally, for the 'ambiguous / unambiguous [Non-binary]' category, we observed errors reaching up to three instances classified as 'error [1]', two as 'error [2]' and one as 'error [3]' (Appendix F). These errors primarily stemmed from difficulties in producing sufficient gender-neutral language. In these cases, the model resorted to using epicene nouns, such as "μηχανικός" (mechanic) and "υπάλληλος" (clerk), while in others, it attempted to use double forms, such as "ο/η συντάκτης/τρια" (the.MASC/FEM editor.MASC/FEM), followed by the plural pronoun "τους" (their). This suggests that the model interpreted "their" as the plural form rather than as a singular non-binary pronoun.

The most interesting results, however, are found in the sentences categorised as 'error [1]', where the model attempted to generate additional gender forms (masculine, female, neuter) as a neutralisation technique. For example, it produced "η/ο/το ρεσεψιονίστ/ρεσεψιονίστρια/ρεσεψιονίστ" (the.FEM/MASC/NEUT receptionist.MASC/FEM/error). Here, "το ρεσεψιονίστ" is a non-existent term in Greek, making it a translation error. Furthermore, in the possessive pronouns at the end of the same sentence, "τη δουλειά του/της/της" (his/her/her work), there is a repetition of the feminine pronoun ("της") instead of a neuter alternative. In an attempt to create gender-neutral language, the model failed to align with grammatical norms.

A similar issue appears in the sentence involving the occupational noun "secretary", where the model produced "ο/η/το γραμματέας" (the.MASC/FEM/NEUT secretary). While "γραμματέας" is an epicene noun used for masculine and feminine references, it does not function as a neuter form, since occupational nouns in Greek lack a neuter equivalent. Additionally, the possessive pronouns in this sentence ("του/της" – his/her) excluded any attempt at a neuter form.

Finally, in the translation classified as 'error [1]', which featured the epicene "συγγραφέας" (writer), the model produced "η συγγραφέας/ο συγγραφέας" (the.FEM writer/the.MASC writer) and introduced the non-existent word "ατους" as the possessive pronoun. The model apparently struggled to produce gender-neutral language leading to mistranslations and non-existent words.

Across these cases, except for those classified as ’error [3]’, it becomes evident that the model aimed to cover different gender possibilities rather than generating a singular gender-neutral form. This strategy is inconsistent with the prompt instructions on how to translate a sentence containing the singular pronoun "they". For three reported cases, GPT-4o produced outputs that included both the masculine and feminine forms. The sentences were grammatically and syntactically correct, however, they are considered non-inclusive as only a gender-neutral translation would be the correct approach when the source purposefully omits gender or refers to a non-binary individual.

Thus, the prompted GPT-4o was overall able to accurately translate and produce gender-inclusive translations for most source sentences (> 77% success rate in each sentence type). These results demonstrate that our hypothesis (H2) is **accepted**, as the model reduced gender bias by providing gendered and gender-neutral options when suitable. That is not uniform, however, as there were examples where the model struggled with non-binary or anti-stereotypical roles. While GPT-4o made significant progress towards inclusivity, there are still other limitations that would need to be further improved, as described in the next sections.

5. DISCUSSION

Our analysis reveals that both Google Translate and DeepL exhibit significant gender bias in English-to-Greek translations, frequently defaulting to masculine forms when the source text does not explicitly indicate a gender (H1a). As reported by prior research on male bias in other language pairs (Prates, Avelar, and Lamb 2019; Stanovsky, Smith, and Zettlemoyer 2019; Currey et al. 2022), we also emphasise how common masculine defaults are in machine translation. We further show that this male bias persists in gender-ambiguous sentences regardless of the stereotypical gender associated with the occupation.

In particular, we observed a significant increase in feminine forms when translating stereotypically female occupations, suggesting that both systems are influenced by gender stereotypes, even while defaulting to masculine forms (H1b). This finding supports previous research (Savoldi et al. 2021) which shows that MT systems reinforce societal gender associations embedded in their training data. It is, thus, reasonable to assume the emergence of a conflict between the male bias and stereotype-driven shifts toward feminine forms.

Moreover, both systems performed consistently well in translating stereotypical and anti-stereotypical gender assignments when the referent's gender was explicitly masculine. As shown in previous studies (Sun et al. 2019; Kocmi, Limisiewicz, and Stanovsky 2020; Saunders and Byrne 2020), translations tend to be more accurate for sentences involving men and for those reflecting stereotypical gender roles (e.g. "male doctor"). When it comes to anti-stereotypical gender assignments involving feminine referents (e.g. "female doctor"), Google Translate struggled more than DeepL. This disparity is likely influenced by differences in training data or model architecture, as the black-box nature of these systems make it difficult to determine the exact cause. DeepL performed better overall, accurately assigning the correct gender in most cases when explicitly defined, though it did fail in one instance.

Lastly, neither system offered gender-neutral or inclusive translations, highlighting their limitation to a binary gender framework. In fact, none of the 'ambiguous / unambiguous [Non-binary]' cases were assigned the neutrality that was required, but instead, in most cases they were assigned binary representations (predominantly masculine, with occasional feminine forms), or led to erroneous translations in which the singular pronoun "they" was misinterpreted as a collective one. The MT systems' inability to handle non-binary or neutral language reflects broader challenges identified in previous research on grammatical gender languages, where binary gender frameworks are deeply ingrained (Savoldi et al. 2021). Overall, these results validate H1, affirming that gender bias persists in machine translations (also) from English into Greek.

Results of the evaluation of the prompted GPT-4o showed that the model performs particularly well in translating gender-unambiguous sentences with masculine forms, while feminine forms were slightly less accurate, suggesting some persistent bias in the training data. Remarkably, DeepL appears to handle feminine gender-unambiguous sentences better than GPT-4o, which suggests that whatever architecture or training data it uses is better able to handle explicit feminine forms. Nevertheless, despite these results, GPT-4o performed far better than both MT systems as it identified and disambiguated gender ambiguities for the majority of ambiguous sentences. The fact that it attempted to generate three alternatives (masculine, feminine, neutral) for all ambiguous cases, as explicitly instructed, indicates that LLMs have the potential to support more inclusive translation practices. These findings align with 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 may underline inherent bias in the training data of the model, but also the challenges in adapting gender-neutral practices for Greek. It is important to consider that gender-neutral and gender-inclusive translation is a challenging and subjective task even for human translators due to a lack of adequate linguistic structures in this language, as was covered in Section 2.2.

6. LIMITATIONS & FUTURE WORK

In addition to the limitations addressed in previous sections, a few more factors should be acknowledged. First, there is a reproducibility problem as the study relies on three closed-source models. Given that the systems are proprietary and regularly updated, the results of the same query may vary across multiple trials. Moreover, GPT-4o requires a paid subscription,[23] which limits its accessibility compared to freely available systems.

Second, GendEL consists of highly controlled sentences, based on certain structures that involve gender-biased occupational nouns and adjectives. This approach provides consistency, however, it restricts the generalisability of the results to more diverse and natural text. In addition, the prompt for GPT-4o is specifically customised for these sentence structures, which raises questions about its applicability to more complex data. GendEL, therefore, should be viewed as the basis for evaluating English-to-Greek translations 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 underrepresented 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

23 \$2.50/1M input tokens and \$10.00/1M output tokens

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) an emphasis on the need for more gender-inclusive translation practices in Greek. We hope that this work will inspire further research on this language pair and contribute to the development of more inclusive translation technologies.

References

Alvanoudi, Angeliki. 2015. *Grammatical gender in interaction: Cultural and cognitive aspects*. Brill.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Cho, Won Ik, 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, Association for Computational Linguistics, Florence, Italy.

Currey, Anna, 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, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

Elaraby, Mostafa, 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.

Escudé Font, Joel 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, Association for Computational Linguistics, Florence, Italy.

Farkas, Anna and Renáta Németh. 2022. How to measure gender bias in machine translation: Real-world oriented machine translators, multiple reference points. *Social Sciences Humanities Open*, 5(1):100239.

Ghosh, Sourojit 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, Association for Computing Machinery, New York, NY, USA.

Gkasouka, Maria and Marianthi Georgalidou. 2014. Οδηγός χρήσης μη-σεξιστικής γλώσσας στα διοικητικά έγγραφα. General Secretariat for Equality and Human Rights, Ministry for Social Cohesion and Family.

Gonen, Hila 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, Association for Computational Linguistics, Online.

Hellenic Open University. 2024. Πρακτικός οδηγός για τη χρήση συμπεριληπτικής ως προς το φύλο γλώσσας στο ΕΑΠ [practical guide for the use of gender-inclusive language at the hellenic open university].

Johnson, Melvin. 2020. A scalable approach to reducing gender bias in google translate.

Kalfadopoulou, Valentini 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.

Karastergiou, Anestis Polychronis and Konstantinos Diamantopoulos. 2024. Gender issues in machine translation. *Transcultural Journal of Humanities Social Sciences*, 5:48–64.

Kocmi, Tom, 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, Association for Computational Linguistics, Online.

Lee, Minwoo, 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, Association for Computational Linguistics, Mexico City, Mexico.

Levy, Shahar, 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, Association for Computational Linguistics, Punta Cana, Dominican Republic.

Měchura, Michal. 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, Association for Computational Linguistics, Seattle, Washington.

Mohamed, Shereen A., Ashraf Elsayed, Yasser Fouad Hassan, and Mohamed Abd-ElRahman Abdou. 2021. Neural machine translation: past, present, and future. *Neural Computing and*

Applications, 33:15919 – 15931.
Mucchi-Faina, Angelica. 2005. Visible or influential? Language reforms and gender (in)equality. *Social Science Information*, 44(1):189–215.
Ntouvlis, Vinicio. 2020. Online writing and linguistic sexism: The use of gender-inclusive @ on a greek feminist facebook page. *Tilburg Papers in Culture Studies*, 245.
Papineni, Kishore, 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, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA.
Pavlidou, Theodossia-Soula, Angeliki Alvanoudi, and Eleni Karafoti. 2004. Grammatical gender and semantic content: Preliminary remarks on the lexical representation of social gender. *Studies in Greek Linguistics*, 24:543–553.
Petikas, Vasilis. 2021. Οδηγός για τη χρήση μη σεξιστικής γλώσσας. Gender Equality Commitee, University of Crete.
Piergentili, Andrea, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023a. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, European Association for Machine Translation, Tampere, Finland.
Piergentili, Andrea, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023b. Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140, Association for Computational Linguistics, Singapore.
Piergentili, Andrea, 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, European Association for Machine Translation (EAMT), Sheffield, UK.
Pino, Marco and David Matthew Edmonds. 2024. Misgendering, cisgenderism and the reproduction of the gender order in social interaction. *Sociology*.
Prates, Marcelo O. R., Pedro H. C. Avelar, and Luis Lamb. 2019. Assessing gender bias in machine translation – a case study with google translate.
Rarrick, Spencer, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 845–854, Association for Computing Machinery, New York, NY, USA.
Rarrick, Spencer, Ranjita Naik, Sundar Poudel, and Vishal Chowdhary. 2024. Gate x-e : A challenge set for gender-fair translations from weakly-gendered languages.
Robinson, Kevin, Sneha Kudugunta, Romina Stella, Sunipa Dev, and Jasmijn Bastings. 2024. Mittens: A dataset for evaluating gender mistranslation.
Rudinger, Rachel, 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, New Orleans, Louisiana.
Sánchez, Eduardo, 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, Association for Computational Linguistics, Miami, Florida, USA.
Saunders, Danielle 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, Association for Computational Linguistics, Online.
Saunders, Danielle, 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, Association for Computational Linguistics, Barcelona, Spain (Online).
Savoldi, Beatrice, 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.
Savoldi, Beatrice, 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, Association for Computational Linguistics, St. Julian's, Malta.

Sennrich, Rico. 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, Association for Computational Linguistics, Valencia, Spain.

Snover, Matthew, 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, Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA.

Stafanovičs, Artūrs, 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, Association for Computational Linguistics, Online.

Stanovsky, Gabriel, 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, Association for Computational Linguistics, Florence, Italy.

Sun, Tony, 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, Association for Computational Linguistics, Florence, Italy.

Triantafyllidis, Manolis. 1963. Η βουλευτίνα και ο σχηματισμός των θηλυκών επαγγελματικών, volume Β. Ίδρυμα Μανόλη Τριανταφυλλίδη.

Troles, Jonas-Dario 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, Association for Computational Linguistics, Online.

Tsokalidou, Roula. 1996. Το Φύλο της Γλώσσας, Οδηγός μη-σεξιστικής γλώσσας για τον δημόσιο ελληνικό λόγο. Σύνδεσμος Ελληνίδων Επιστημόνων-Βιβλιοπωλείο της Εστίας.

Tsopanakis, Agapitos. 1982. Ο δρόμος προς την δημοτική: Θεωρητικά, τεχνικά και γλωσσικά προβλήματα. Σχηματισμός επαγγελματικών θηλυκών. In Ὁ δρόμος πρός τήν Δημοτική (Μελέτες και ἄρθρα). Εκδοτικός οίκος Αφών Κυριακίδη, pages 302–342.

Vanmassenhove, Eva. 2024. Gender bias in machine translation and the era of large language models. *ArXiv*, abs/2401.10016.

Vanmassenhove, Eva, Chris Emmery, 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, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Vanmassenhove, Eva, 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, Association for Computational Linguistics, Brussels, Belgium.

Wang, Boshi, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Association for Computational Linguistics, Toronto, Canada.

Zhao, Jieyu, 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, New Orleans, Louisiana.

Zmigrod, Ran, 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, Association for Computational Linguistics, Florence, Italy.

Appendix A: 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 δικηγορίνα** 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 αναξιόπιστος δικηγόρος** τελείωσε τη δουλειά.	**Η αναξιόπιστη δικηγορίνα** τελείωσε τη δουλειά.	**Το αναξιόπιστο 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 δικηγορίνα** 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 δικηγορίνα** 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.

Appendix B: 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}

Appendix C: Absolute Count Tables

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

Table 1
Absolute counts for gender categories and errors in translations by Google Translate and DeepL. This table complements Figure 5.

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

Table 2
Absolute counts for gender distribution of translations for **stereotypically male occupations** in **gender-ambiguous sentences**, produced by Google Translate and DeepL. This table complements Figure 6.

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

Table 3
Absolute counts for gender distribution of translations for **stereotypically female occupations** in **gender-ambiguous sentences**, produced by Google Translate and DeepL. This table complements Figure 7.

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

Table 4
Absolute counts for gender distribution of translations for **stereotypically male occupations** in **masculine gender-unambiguous sentences** (stereotypical case), produced by Google Translate and DeepL. This table complements Figure 8.

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

Table 5
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. This table complements Figure 9.

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

Table 6
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. This table complements Figure 10.

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

Table 7
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. This table complements Figure 11.

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

Table 8
Absolute counts for gender distribution of translations for **masculine gender-unambiguous sentences** produced by Google Translate, DeepL and prompted GPT-4o. This table complements Figure 13.

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

Table 9
Absolute counts for gender distribution of translations for **feminine gender-unambiguous sentences** produced by Google Translate, DeepL and prompted GPT-4o. This table complements Figure 14.

Appendix D: Results of Fischer's Exact Test

Metric	Google Translate	DeepL
Odds ratio	0.0	0.0
p-value	0.000143	2.105037e-07

Table 1
Results of Fischer'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
p-value	1.0	inf

Table 2
Results of Fischer's exact test for Google Translate and DeepL, investigating whether there is a statistically significant 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
p-value	0.018701	n/a

Table 3
Results of Fischer's exact test for Google Translate, investigating whether there is a statistically significant 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.

Appendix E: Gender Distribution of Prompted GPT-4o per Sentence Type

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 1
Absolute counts and percentages for gender categories and errors in translations by prompted GPT-4o. This table complements Figure 12.

Appendix F: Qualitative Analysis of Errors 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 1
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 2
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 3
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 4
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 5
Error instances produced by prompted GPT-4o for the ʹambiguous / unambiguous [Non-binary]ʹ sentences.