Assumed Identities: Quantifying Gender Bias in Machine Translation of Gender-Ambiguous Occupational Terms

Orfeas Menis Mastromichalakis[1], **Giorgos Filandrianos**[1,2], **Maria Symeonaki**[3] and **Giorgos Stamou**[1]
[1]School of Electrical and Computer Engineering, National Technical University of Athens, Greece
[2]Instituto de Telecomunicações, Portugal
[3]Department of Social Policy, Panteion University of Social and Political Sciences, Greece
{menorf,geofila}@ails.ece.ntua.gr, msymeon@panteion.gr, gstam@cs.ntua.gr

Abstract

Machine Translation (MT) systems frequently encounter gender-ambiguous occupational terms, where they must assign gender without explicit contextual cues. While individual translations in such cases may not be inherently biased, systematic patterns—such as consistently translating certain professions with specific genders—can emerge, reflecting and perpetuating societal stereotypes. This ambiguity challenges traditional instance-level single-answer evaluation approaches, as no single gold standard translation exists. To address this, we introduce GRAPE, a probability-based metric designed to evaluate gender bias by analyzing aggregated model responses. Alongside this, we present GAMBIT, a benchmarking dataset in English with gender-ambiguous occupational terms. Using GRAPE, we evaluate several MT systems and examine whether their gendered translations in Greek and French align with or diverge from societal stereotypes, real-world occupational gender distributions, and normative standards[1].

1 Introduction

Machine Translation systems have become indispensable tools for cross-linguistic communication, yet they frequently exhibit gender biases that reinforce societal stereotypes (Blodgett et al., 2020; Menis-Mastromichalakis et al., 2025). In the labour market, where gender disparities persist, such biases are particularly concerning. For example, as illustrated in Figure 1, Google Translate[2] systematically assigns masculine grammatical forms to occupations traditionally dominated by men or stereotypically perceived as masculine (e.g., CEO, doctor, plumber), and feminine forms to those commonly associated with women (e.g., nurse, secretary, cleaner) when translating gender-ambiguous inputs from English to Greek. This is not an isolated case, but a consistent pattern across most MT systems (Alvarez-Melis and Jaakkola, 2017; Escudé Font and Costa-jussà, 2019). These biases extend beyond language, subtly validating and reinforcing occupational segregation by shaping perceptions of gender roles. This, in turn, influences hiring practices, career aspirations, and wage disparities, further entrenching systemic inequalities in the workforce (European Commission, 2020). Addressing these biases is essential to ensure that MT systems contribute to fair representations of professions rather than perpetuating historical and cultural stereotypes.

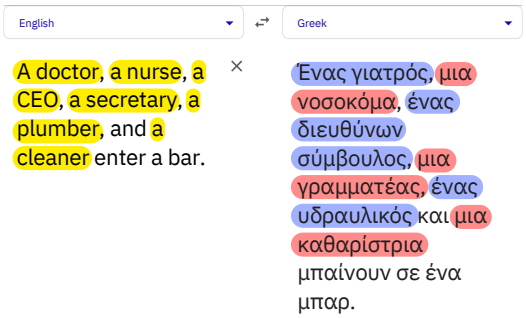

Figure 1: An example of gender stereotypes reflected in a translation from English to Greek (Google Translate). All occupations in the source text (highlighted in yellow) are gender-ambiguous, while target terms highlighted in blue indicate masculine grammatical forms and those in red indicate feminine forms.

Evaluating occupational gender bias in MT systems is particularly challenging when the occupational terms are gender-ambiguous. When translating from genderless (e.g., Finnish or Turkish) or notional gendered languages (e.g., English) into languages with grammatical gender (e.g., Greek or French), MT systems often must make assumptions and assign gender, as preserving ambiguity

[1]Our code is available at https://github.com/ails-lab/assumed-identities, and the GAMBIT dataset is publicly available at https://huggingface.co/datasets/ailsntua/GAMBIT.

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

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32222–32238
November 4-9, 2025

or using gender-neutral language is not always feasible or stylistically appropriate. Unlike explicit biases, such as misgendering occupations with clear gender markers in the input, which can be directly flagged as incorrect, these cases exhibit a unique duality: a single translation—e.g., translating "the actor" as *"l'acteur"* (masculine) or *"l'actrice'* (feminine) in French—is not biased or unbiased in isolation, as both choices are equally valid in the absence of contextual cues. However, when examined in aggregate, systematic patterns may emerge, revealing a model's predisposition to associate certain professions with specific genders. This renders traditional instance-level single-answer quality (Papineni et al., 2002; Lin, 2004) or bias (Stanovsky et al., 2019) evaluation approaches unsuitable, as they fail to capture these broader distributional trends in the absence of a gold standard.

In this work, we shift the focus from isolated translations to aggregated model behavior, enabling the detection of systematic gender biases that may not be evident at the individual sentence level. We propose a methodology to detect, classify, and quantify gender assignments in the translation of gender-ambiguous occupational terms. To support this evaluation, we introduce GAMBIT (GenderAMBIguous occupaTions), a benchmarking dataset of English texts containing occupational terms expressed in a gender-neutral or ambiguous way. Our approach identifies gender assignments by comparing source texts with their translations and aggregates gendered outputs across multiple instances. Occupations are grouped using the International Standard Classification of Occupations (ISCO-08)[3], to enable analysis at varying levels of abstraction and account for lexical variations. ISCO-08 is an internationally recognized system for classifying occupations endorsed by the International Labour Organisation (ILO). It provides a hierarchical structure that categorizes jobs into four levels of increasing granularity, using a digit-based coding system. At the highest level, occupations are grouped into broad categories, which are then divided into more specific subcategories at lower levels. For example, the top-level category "Professionals" (code 2) includes subcategories such as "Science and Engineering Professionals" (code 21) and "Health Professionals" (code 22), that further divide into "Medical Doctors" (code 221), "Nursing and Midwifery Professionals" (code 222), and others. Each category in ISCO-08 is accompanied by detailed descriptions, examples of occupations, and other relevant information, providing a comprehensive framework for analyzing and comparing jobs across different countries and industries. GAMBIT spans the entire ISCO-08 taxonomy, ensuring broad occupational coverage. To quantify bias, we introduce GRAPE, a probability-based metric that measures divergence from reference distributions, such as idealized gender parity or real-world labor statistics. We apply our framework to translations from English into Greek and French, two languages with grammatical gender but from different language families, comparing outputs against both normative standards and empirical labor data. Our approach offers a scalable and interpretable framework to evaluate gender bias in MT, offering insights into how translation systems reflect, reinforce, or potentially challenge societal patterns of occupational gender representation when facing ambiguity.

[3] https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/

2 Related Work

Research on gender bias in NLP has explored a broad range of tasks and provided valuable insights (Bolukbasi et al., 2016; Lu et al., 2020), but our focus is on Machine Translation (Savoldi et al., 2021; Vanmassenhove, 2024), where gender bias remains a pressing issue with significant societal impact (Savoldi et al., 2024). Numerous case studies have highlighted the prevalence and consequences of gender bias in MT across languages and cultural contexts (Rescigno et al., 2020; Farkas and Németh, 2022; Ghosh and Caliskan, 2023; Paolucci et al., 2023; Kostikova et al., 2023; Piazzolla et al., 2023), emphasizing the need for effective evaluation and mitigation. Moreover, critiques of existing quality metrics reveal that traditional evaluation methods often fail to capture gender disparities adequately (Zaranis et al., 2024). To tackle this, researchers have developed resources and methods that target gender bias in MT. This includes Knowledge Graphs that offer structured, contextual information for bias analysis (Mastromichalakis et al., 2024), multilingual benchmarks (Currey et al., 2022), and studies on language-specific challenges such as gender-neutral pronouns (Cho et al., 2019). Alongside, mitigation efforts (Sun et al., 2019) explore model fine-tuning, data balancing, and adaptive learning (Escudé Font and Costa-jussà, 2019; Saun-

ders and Byrne, 2020; Costa-jussà and de Jorge, 2020), with recent work also focusing on gender-neutral and gender inclusive translation strategies (Piergentili et al., 2023a; Lardelli and Gromann, 2023) and benchmarking such approaches (Piergentili et al., 2023b; Lardelli et al., 2024; Gkovedarou et al., 2025).

Our work studies occupational gender bias in translating gender-ambiguous inputs, adding to ongoing research on gender bias in NLP with a focus on occupations and the labor market (Tal et al., 2022; Gorti et al., 2024). Ambiguity has also been studied in other NLP tasks, such as Question Answering (Li et al., 2020; Parrish et al., 2022) and coreference resolution (Rudinger et al., 2018; Zhao et al., 2018), where multiple plausible interpretations reveal the influence of stereotypes. For example, Kotek et al. (2023) examine ambiguous coreference inputs in LLMs, where no single ground truth exists. Their aggregated analysis of role, trait, and occupation associations exposes stereotypical patterns, aligning with our approach of studying model behavior at an aggregated level to detect subtle biases. In MT, one way to handle ambiguity is by generating all grammatically correct gendered translations (Garg et al., 2024), a strategy used by some commercial systems. While inclusive, this approach is limited to setups that allow multiple outputs and faces scalability challenges as multiple ambiguities exponentially increase possible translations. Other approaches disambiguate inputs before translation (Vanmassenhove et al., 2018), which however requires some structural or semantic hints that allow the disambiguation of gender. This is the case for some challenge sets like WinoMT (Stanovsky et al., 2019) where gender can be disambiguated via correference, or the MuST-SHE corpus (Bentivogli et al., 2020) that includes audios and transcripts, where the inputs have a correct gender resolution due to gender cues that are recoverable from audio (e.g., speaker's voice) or textual context (e.g., pronouns, named entities).

In contrast, our study focuses on ambiguous cases without disambiguating cues, allowing inherent stereotypical associations and biases to emerge naturally. Our inputs are deliberately designed to have multiple plausible interpretations without a single correct answer. Gender ambiguity in MT has been explored through a range of challenge sets and benchmarks, yet most existing efforts remain limited in scope, scale, or evaluation depth. gENder-IT (Vanmassenhove and Monti, 2021) introduced a manually curated English–Italian challenge set stemming from MuST-SHE, covering natural gender phenomena, including occupation-related examples. While it includes truly ambiguous instances, the dataset remains limited in scale (694 sentences), treats each sentence in isolation, and lacks a structured evaluation methodology. A follow-up study (Vanmassenhove, 2024) used a subset of gENder-IT to assess ChatGPT's performance, but provided only a brief and high-level analysis. Concurrent to our work, Hackenbuchner et al. (2025) introduced GENDEROUS, a handcrafted dataset of sentences with statistically stereotypical occupational nouns and gender-inflected adjectives, and examined the effect of gender-ambiguous inputs. Other works have explored bias through contrastive sentence pairs. Gonen and Webster (2020) for instance, generate minimal pairs differing by a single human-related noun to expose gender asymmetries in translations. In a different approach, Prates et al. (2020) examine gender bias in Google Translate using simple templates. While their aggregated analysis shares similarities with ours, the reliance on templated inputs and the focus on a single MT system restricts the generalizability of their findings. Our work on the other hand, introduces a comprehensive evaluation framework that goes beyond challenge sets. It covers a broad range of occupations based on ISCO classifications, provides rich contextual texts rather than isolated sentences, enables structured, multilingual evaluation through interpretable, statistics-informed metrics, and evaluates a variety of MT systems.

3 Detecting Gender Assignments

Our approach focuses on detecting gender assignments in the translation of occupational terms when the source gender is ambiguous. By gender assignment, we refer to cases where the translation introduces a masculine or feminine form not specified in the source. Given a gender-ambiguous input, we feed it to an MT system and analyze the output to determine the gender of any translated occupational terms. This process follows an approach inspired by the "LLM-as-a-judge" paradigm (Li et al., 2025; Gu et al., 2025), a framework that has previously been applied to evaluate gender bias in machine translation systems and textual corpora (Derner et al., 2024; Piergentili et al., 2025), and consists of two main steps: (1) detecting occupations in the translation, and (2) identifying their gender.

For the first step, we employ an LLM-based component in a few-shot setup, prompting it to extract all explicitly mentioned occupation titles in the text. During development, we identified two main types of hallucinations and designed targeted strategies to mitigate them. The first type occurs when the LLM *identifies occupations that are not present in the text*. For example, in the sentence: "The supplier complained to the call center," the LLM may incorrectly detect the occupation "Customer Service Representative", even though it is not explicitly mentioned. Note that this example is given in English for illustrative purposes; in practice, such cases only arose in the French and Greek translations, since occupation and gender detection were not performed on the English source texts (which were already provided by the GAMBIT dataset). To address this, we instructed the LLM to provide both the detected occupation titles and their corresponding in-text occurrences. We then applied fuzzy string matching to verify whether the detected terms appeared in the text. If the similarity fell below a predefined threshold, the term was discarded as a hallucination. The second type of hallucination involves the LLM *incorrectly identifying non-occupational terms as occupations*. This issue was particularly prevalent in cases where no occupations were present in the input text. For instance, in "She is a master in her craft.", the word "Master" was wrongly detected as an occupation. To address this, we modified the prompt to require the LLM to also generate a short description for each detected occupation. This description serves as a verification step to check whether the detected occupation matches any ISCO-08 entry. To do this comparison, we used an embedding-based approach. We converted both the LLM-generated descriptions and the ISCO-08 descriptions into embeddings and applied cosine similarity to find the closest match. Following the method from Li and Li (2024), we used angle-based embeddings to map the descriptions into a common latent space. If the similarity score between the LLM's description and any ISCO-08 occupation was below a threshold, the detected occupation was discarded. This step improved both accuracy and consistency by ensuring alignment with ISCO-08 occupations. In our case, the comparison was limited to a known set of candidate occupations from the source text, which simplified the task and made thresholding more efficient.

The second step of our approach involves identifying the gender of the detected occupations, assigning to them one of the three labels: *"Masculine,"* *"Feminine,"* or *"Not Clear"*. This is done using the same LLM, within the same session. After detecting an occupation, the LLM is prompted to assign a gender label to each identified occupation.

This pipeline allows us to detect and measure gender assignments between source and translated texts. These assignments are then aggregated using the evaluation framework described in Section 4 to quantify the model's gender bias. Technical implementation details are provided in Appendix A, and all prompts are listed in Appendix B.

4 Evaluation Framework

In this section, we present an evaluation framework to study occupational gender bias in Machine Translation systems when handling gender-ambiguous inputs. Our goal is to quantify gender bias by analyzing the distribution of gendered translations across these ambiguous cases, revealing patterns of bias that instance-level evaluations may overlook.

In real-world applications, MT systems typically produce a single output, forcing a choice when ambiguity is present. In these cases, the system makes an implicit assumption, raising the question of how this decision should be evaluated. Here, two, sometimes competing, perspectives emerge: *normative correctness* and *predictive accuracy* (Deery and Bailey, 2022). Normative correctness evaluates system behavior against an idealized standard of fairness, such as gender parity. Predictive accuracy, on the other hand, assesses how well the system reflects a reference distribution, such as real-world gender statistics for a given occupation.

Since our approach aggregates behavior across multiple outputs and we do not expect consistent behavior across all occupations, it is essential to group outputs that refer to the same occupation(s). This will allow a more detailed investigation, identifying the model's associations between specific occupations and gender. However, simple keyword-based clustering is inadequate due to variation in how occupations are expressed, and could lead to fragmented or inconsistent clusters. Moreover, many occupations are semantically related, while others differ significantly. Analyzing each occupation in isolation limits the ability to draw generalizable conclusions.

To address this, we adopt ISCO-08 as our oc-

cupation taxonomy. It allows us to cluster, organize, and analyze translations at multiple levels of abstraction, beginning with detailed (4-digit) categories. This enables us to investigate gender bias both at the level of specific occupations and across broader occupational groupings. Beyond clustering, ISCO-08 also captures hierarchical relationships among occupations, which we use to structure our analysis and identify patterns of bias that span related roles. Using ISCO-08 ensures consistency and comparability across occupations, supports structured and meaningful generalization, and aligns our framework with international standards. This, in turn, facilitates comparisons with real-world labor statistics and improves the interpretability and applicability of our findings.

4.1 Metrics

To quantify gender bias in MT and compare model outputs against reference distributions, we introduce the *Gender RAtio ProbabilitiEs* (GRAPE). This metric measures how the likelihood of generating masculine or feminine forms for gender-ambiguous terms diverges from a chosen reference distribution.

Definition 1 (Gender RAtio ProbabilitiEs (GRAPE))**.** *Let M be a set of source–target text pairs, where each source contains a gender-ambiguous term. Let p_m be the observed probability of generating a masculine form in M, and $p_f = 1 - p_m$ the probability of generating a feminine form. Let p_m^{ref} denote the reference probability for the masculine form, and $p_f^{ref} = 1 - p_m^{ref}$ for the feminine.*
GRAPE is defined as:

$$GRAPE_g^{ref}(M) = \frac{p_g - p_g^{ref}}{p_g^{ref}}, g \in \{m, f\} \quad (1)$$

Positive values indicate bias toward the respective gender, while negative values against it.

Intuitively, GRAPE measures the relative difference between the model's output probability for a gendered form and the corresponding reference probability. For example, $\text{GRAPE}_{\text{m}}^{\text{ref}}(M) = 1.0$ implies that the system generates masculine forms twice (100%) more often than expected based on the reference. These metrics quantify both the *direction* and *magnitude* of gender bias.

Although MT outputs are not always strictly binary in gender, maintaining ambiguity is often impractical in gendered languages due to grammatical and morphological constraints. Some languages, such as Greek, include epicene occupational terms that are identical for masculine and feminine forms. However, even in these cases, gendered pronouns and articles often reveal gender. Additionally, historically masculine epicene terms (e.g., βουλευτής) are increasingly complemented by feminine forms (e.g., βουλεύτρια), reflecting evolving usage in discourse and media. These factors make it difficult for MT systems to preserve gender ambiguity in translation. While some systems use gender-neutral strategies (e.g., they/them in English), such approaches are not yet widespread or standardized. In our evaluation (Section 5), gender ambiguity was preserved in fewer than 15% of instances on average. Furthermore, reference distributions (e.g., parity or real-world statistics) typically lack a neutral category, making it difficult to include gender-neutral outputs in our framework. We therefore focus on binary gender forms and leave the integration of neutrality to future work.

The choice of reference distribution is central to interpreting the metrics, and we adopt two perspectives:

- **Normative Correctness**: Assumes ideal gender parity by setting $p_{\text{m}}^{\text{ref}} = p_{\text{f}}^{\text{ref}} = 0.5$. This baseline reflects an expectation of equal representation. In this case we use *ref*=parity.
- **Predictive Accuracy**: Uses empirical data (e.g., labor statistics) to reflect actual gender distributions across occupations. This enables contextual evaluation grounded in real-world demographics. In this case we use *ref*=real.

By applying both perspectives, our evaluation framework captures different dimensions of fairness: one based on equality, the other on realistic alignment.

4.2 Benchmarking Dataset

To enable a comprehensive evaluation of MT systems across the full range of occupations in the ISCO taxonomy, we created GAMBIT, a benchmarking dataset containing English texts with gender-ambiguous occupational terms. Existing datasets (Rudinger et al., 2018; Zhao et al., 2018; Stanovsky et al., 2019) lacked sufficient occupational coverage, particularly in gender-ambiguous contexts, so we opted to generate the dataset using large language models (LLMs), followed by thorough manual review for quality assurance. All

generated instances were validated by domain experts, namely PhD holders in gender studies, social policy, and sociology, with established expertise in occupation-related topics. Validation involved discarding any texts that were not fluent or where occupational terms were not gender-ambiguous. The experts carried out this task as part of their work in a funded project and were compensated according to national standards. Before generating GAMBIT, we attempted to build a dataset from real-world data by processing over 250,000 random texts from the WMT[4] and C4[5] (Dodge et al., 2021) datasets. However, this approach yielded data for only 43 ISCO unit groups (less than 10% of the 436 total), with limited textual diversity and repetitive patterns that could introduce bias. This made artificial generation the only viable approach for ensuring both full occupational coverage and a variety of textual styles.

For the generation, we used Claude 3.5 Sonnet[6]. Detailed information about the prompts is provided in Appendix B. We collected all occupational titles from each 4-digit ISCO-08 class and generated multiple examples per occupation, varying by text format. GAMBIT consists of 9,805 English samples, averaging 22.5 texts per occupation, distributed evenly across five formats: short stories, brief news reports, short statements, short conversations, and short presentations (1,961 samples per format). Detailed statistics on character and word length are provided in Appendix C. The dataset is designed to support gender bias evaluation for any language pair with English as the source language. Adapting the methodology proposed in Section 3 to a different target language requires only minimal changes, as the core detection components rely on LLMs, which are available for most languages.

4.3 Real World Statistics

To calculate the reference distribution in the predictive approach of our metrics, we collected real-world labor statistics. Specifically, we collected the gender-based occupational distributions for France and Greece, since our translation tasks involve English to French and English to Greek, respectively. Although both languages are also spoken in other parts of the world, we focused on these countries as representative examples to demonstrate how our approach can incorporate real-world demographics.

We analyzed raw microdata drawn from the European Union Labour Force Survey (EU-LFS)[7]. This large-scale sample survey provides quarterly and annual statistics on labor participation and inactivity among individuals aged 15 and older, using standardized definitions and the ISCO-08 classification to ensure cross-country comparability. In particular, we calculated the gendered occupational distributions at the ISCO-08 3-digit level for both countries over the period 2011-2023. This allows us to benchmark MT systems against real-world occupational gender distributions, providing a meaningful reference point for evaluating gender bias. As these statistics follow the ISCO-08 classification, they align directly with our benchmark dataset, enabling straightforward mapping between the two.

5 Experiments and Results

5.1 Pipeline's Performance

To evaluate the performance of the pipeline introduced in Section 3, we constructed a separate validation dataset in French and Greek. Existing datasets were either limited to English or covered only a narrow range of occupations. Therefore, we followed the same construction approach as for GAMBIT (see 4.2), ensuring broad occupational coverage across the ISCO classification and textual variety. The datasets included masculine, feminine, and, where possible, gender-neutral forms of occupations, allowing us to assess how well the pipeline handles gender-specific and ambiguous cases. Each language dataset contains 29,415 texts, with occupations evenly distributed across the ISCO taxonomy. A random sample of approximately 20% of the data was manually reviewed, and no issues with the text or labels were found. Further details are provided in Appendix D.

Table 1 presents the pipeline's performance, reporting accuracy in identifying occupations, detecting gender, and combining both, using Claude 3.5 Sonnet. The results show that the pipeline reliably extracts both occupation and gender information, making it a suitable tool for analyzing gender-related behavior in machine translation systems.

5.2 Analysis of MT systems

We evaluated several widely used MT systems, including Google Translate, M2M100 (Fan et al., 2021), and NLLB (Costa-Jussà et al., 2022) with

[4] https://huggingface.co/datasets/wmt/wmt14
[5] https://huggingface.co/datasets/allenai/c4
[6] Model ID: anthropic.claude-3-5-sonnet-20241022-v2:0
[7] https://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey

Lang	Occ. Acc.	Gender Acc.	Overall
French	99.93	98.30	98.30
Greek	99.92	99.53	99.47

Table 1: Pipeline accuracy

600 million and 1.2 billion parameters, as well as LLMs like Claude-3.5, and EuroLLM (Martins et al., 2024), prompted to perform translations. Further details on the models and implementation are provided in Appendix A and the prompts used for the LLM translations in Appendix B.

5.2.1 Overall Behavior

We first examined the overall behavior of the MT systems, focusing on how biased their outputs are, and how they align with gender parity and real-world occupational distributions. Table 2 presents GRAPE for masculine and feminine translations across systems, using both ideal parity and real-world data as reference points. The results show a clear and consistent trend: MT systems overwhelmingly translate gender-ambiguous texts into masculine forms. This confirms previous findings that MT models often adopt a *masculine as default* strategy when gender is unclear (Schiebinger, 2014; Vanmassenhove et al., 2018; Monti, 2020). Notably, this tendency is not unique to automated systems. It reflects broader patterns in human language use, where masculine forms are commonly used in situations of gender ambiguity, not only in translation but also in everyday communication (Silveira, 1980; Stahlberg et al., 2011). This is likely reflected in the training data used for MT systems and LLMs, leading to this bias towards masculine forms.

A tendency toward extreme gendering is also evident at the level of individual occupations, however not always towards masculine forms. On average, in 374 out of 436 ISCO occupations (4-digit level), the systems assigned one gender in more than 80% of the translated texts, with the percentage being 100% (i.e. $GRAPE_m^{parity} = 1$ or $GRAPE_f^{parity} = 1$) in 293 of them. Only about 30 occupations showed a more balanced output, with gender assignments falling between 30–70% (see Appendix E for per-model breakdowns). While this confirms the dominant masculine bias, as the vast majority of the extreme cases were masculine-dominated, it also points to a broader issue: models tend to rigidly associate specific occupations with specific genders. In most cases, variation in format, type, and context of the input texts had little effect on the gendered output. This may reflect an underlying tendency of current MT systems to reinforce strong associations learned during training—especially when the task permits or encourages confident, consistent outputs. While such determinism can be useful in many settings, it may also limit the model's ability to reflect ambiguity or diversity.

5.2.2 Influence of Gender Stereotypes

Despite the overall masculine skew, this tendency is not uniform across all occupations. In fact, all models consistently translate a small number of occupations predominantly into feminine forms. The occupations translated into feminine forms are largely consistent across all MT systems and both languages. These include stereotypically feminine roles such as 'Midwifery Professionals' (2222) and associate professionals (3222), 'Nursing Professionals' (2221), and 'Cleaning and Housekeeping Supervisors in Offices, Hotels, and Other Establishments' (5151). In contrast, the occupations translated into masculine forms include not only stereotypically masculine roles—such as miners (8111), house builders (7111), and judges (2612)—but also those perceived as gender-neutral, like visual artists (2651) and high school teachers (2330).

This suggests that MT systems tend to translate stereotypically feminine occupations into feminine forms, while defaulting to masculine for both stereotypically masculine and neutral roles. To validate this, we analyzed $GRAPE_m^{parity}$ across occupations categorized by gender stereotypes as masculine, feminine, and neutral. While real-world gender distributions are often used as proxies for stereotypes, they are not entirely aligned. Research shows that occupational gender stereotypes may reflect outdated perceptions rather than current workforce statistics, with notable mismatches in certain roles (Gygax et al., 2016). To assess stereotypical perceptions directly, we used ratings from Shinar (1975), who provide stereotype scores (1 to 7) for 129 occupations. Appendix F details how we processed this data to group occupations by perceived gender. Using these groupings, we calculated $GRAPE_m^{parity}$ for all models in both languages. The results, shown in Table 3, confirm our observations: all systems predominantly use masculine forms for stereotypically masculine and neutral occupations, while showing more balanced or feminine-leaning translations for stereotypically

MT	*ref*=parity				*ref*=real			
	French		Greek		French		Greek	
	m	f	m	f	m	f	m	f
NLLB-600M	0.92	-0.92	0.91	-0.91	0.88	-0.91	0.67	-0.90
NLLB-1.3B	0.66	-0.66	0.58	-0.58	0.63	-0.65	0.38	-0.51
M2M100	**0.94**	**-0.94**	0.95	-0.95	**0.90**	**-0.94**	0.71	-0.95
EuroLLM-1.7B	0.86	-0.86	0.78	-0.78	0.82	-0.85	0.56	-0.75
GT	0.92	-0.92	**0.97**	**-0.97**	0.89	-0.92	**0.72**	**-0.97**
Claude	0.87	-0.87	0.92	-0.92	0.83	-0.86	0.67	-0.90

Table 2: GRAPE calculated on the whole GAMBIT dataset for the two genders across the different MT systems used in the study. Highest absolute values are depicted in **bold**, while the lowest are underlined.

feminine occupations.

5.2.3 Divergence from the Real World

Our findings show that MT systems do not simply reflect real-world gender imbalances—they often amplify or even distort them. While gender gaps in certain occupations still exist, the models tend to exaggerate these differences or, in some cases, completely reverse them. For instance, most systems translated texts related to 'Administrative and Specialised Secretaries' (ISCO code 334) predominantly into masculine forms in French—over 80% of the time—despite the fact that in 2023, more than 90% of people in this occupation in France were women. Additionally, as shown in Table 2 most systems produce masculine forms nearly twice as often as what real-world statistics suggest.

While true gender equality in the labor market is still far from reality, recent data shows clear progress in reducing gender segregation across occupations. Women today participate in a much broader range of professions than in the past, and the overall numbers of employed men and women are approaching balance in many countries. However, the behavior of MT systems does not reflect this progress. Instead, their outputs often resemble labor patterns from decades ago, when women were largely confined to a limited set of roles such as nurses, or cleaners. This means that even if the models themselves are not getting worse, they diverge more and more over time from the real world because society moves forward, while the systems remain stuck in outdated patterns. As a result, the gap between model outputs and present-day labor realities slowly grows.

To better understand what shapes model behavior, we compared how closely the model outputs align with gender stereotypes versus real-world labor statistics (see Appendix G). We found that the correlation with stereotypical perceptions is slightly—but consistently—higher than with actual employment data. Stereotypes often reflect outdated or oversimplified views of gender roles, and their influence on model behavior points to deeper biases in the underlying datasets. As widely acknowledged in the literature (Leavy et al., 2020; Bender et al., 2021), training data frequently underrepresents female, minority, and non-Western perspectives, while favoring sources that reinforce dominant norms. These imbalances in representation—and in how information is structured—can amplify stereotypical associations. Importantly, even if training data were to perfectly mirror present-day labor statistics, models might still form overly rigid associations, such as consistently linking certain jobs with one gender. This highlights that data alone is insufficient to prevent biased behavior; model architecture, training objectives, and design decisions also play a crucial role.

5.2.4 Bias Alignment

To examine whether gender biases in MT systems are shared across languages, we computed the correlation of gendered translation distributions between the two target languages for each model. Most models showed strong cross-lingual correlations (mean $r = 0.757 \pm 0.140$), indicating that their gender biases are largely consistent across languages. This may suggest that many systems may rely on a shared internal representation that transfers similar gender preferences across languages, or simply that language and people share common gender biases across languages. NLLB-1.3B exhibited a notably lower correlation ($r = 0.478$), which aligns with its overall lower bias and reduced preference for masculine defaults as indicated in Table 2. This may indicate a more language-specific approach to gender, rather than a shared cross-lingual bias. Additionally, NLLB-1.3B showed consistently lower alignment with other models across individual languages, while the remaining models were more similar to each other. Full correlation

MT	masculine		neutral		feminine	
	French	Greek	French	Greek	French	Greek
NLLB-600M	0.96	0.94	0.94	0.94	0.07	0.14
NLLB-1.3B	0.79	0.75	0.67	0.72	-0.19	0.14
M2M100	0.95	0.97	0.97	0.95	0.42	0.34
EuroLLM-1.7B	0.95	0.92	0.87	0.78	-0.09	-0.22
GT	0.96	0.99	0.95	0.99	0.09	0.46
Claude	0.95	0.97	0.86	0.95	-0.23	-0.04

Table 3: $GRAPE_m^{parity}$ for stereotypically masculine, neutral, and feminine occupations.

scores are provided in Appendix H.

6 Conclusions

In this work, we explored the evaluation of MT systems when translating gender-ambiguous occupational terms. We introduced a pipeline to detect gender assignments as an indicator of potential gender bias and proposed a probability-based metric to quantify this bias against reference distributions. This approach allows for evaluation against normative standards, such as equal gender representation, as well as real-world distributions. Additionally, we provided a comprehensive benchmarking dataset containing nearly 10,000 English texts with gender-ambiguous occupational terms, covering the entire ISCO-08 spectrum of occupations. Using this framework, we evaluated 6 widely used MT systems with diverse characteristics, demonstrating the valuable insights our approach can provide.

Future work will focus on adapting the methodology to un-annotated texts, enabling gender bias evaluation of MT datasets and expanding the analysis to more languages. Furthermore, we aim to expand our framework to be able to evaluate gender-neutral translations as well, aligning with current efforts in the field to promote inclusivity and responsible language generation.

Acknowledgments

We want to thank Markella Challiori for introducing us to the debate on normative correctness versus predictive accuracy, which informed the framing of this work, and for her valuable insights during our discussions on gender biases in AI models.

This work was carried out within the framework of the Pharos AI Factory project, funded by the European High-Performance Computing Joint Undertaking (EuroHPC JU) under Grant Agreement No. 101234269 as part of the Horizon Europe and by the Greek Public Investments Program programme.

This work was supported by the FCT project "OptiGov", ref. 2024.07385.IACDC (DOI 10.54499/2024.07385.IACDC), funded by the PRR under the measure RE-C05-i08.m04.

Limitations

A limitation of our work is the use of AI systems that may themselves be biased to detect gender biases in MT. However, we employ these systems for more narrowly defined tasks—namely, occupation detection and gender attribution—that are comparatively simpler and less ambiguous than the overall MT task being evaluated. This focused application reduces the likelihood of the models' inherent biases significantly impacting our results, as also evidenced by our method's near-perfect accuracy in occupation and gender detection. Additionally, it is worth noting that many state-of-the-art MT evaluation metrics, such as COMET (Rei et al., 2020), are themselves based on large language models, which further supports the suitability of LLMs for evaluating translation quality and related properties. This alignment with established practices underscores the reliability of using LLMs in our evaluation framework.

Furthermore, a limitation of our work is the use of an artificially created dataset, which, while properly curated and manually inspected, may still carry some inherent constraints. We acknowledge that relying on such a dataset could introduce biases or limitations in terms of its representation of real-world data. However, this was the only viable option, as no existing dataset with the necessary characteristics for our study—specifically one that includes gender-ambiguous occupational terms across a wide range of occupations—was available. Despite this, the careful curation and expert review of the dataset aimed to minimize potential issues and ensure its reliability for the purpose of our analysis.

Another limitation is our treatment of gender as a binary feature, despite the growing recognition of gender as a spectrum as well as technical approaches to gender-neutral translations. From

a grammatical perspective, our classification into masculine, feminine, and "not clear" partially addresses this complexity to some extent. However, this binary approach remains an oversimplification that fails to capture the full diversity of gender identities, which could be potentially harmful. Nonetheless, real-world statistics are predominantly published using a binary gender framework, making it challenging to analyze this issue in a more nuanced way.

Lastly, as societal roles evolve, so do occupations. Emerging professions, such as content creator or prompt engineer, may not be adequately represented in the ISCO-08 classification and thus are not fully captured in our analysis. In future work, we aim to incorporate these "emerging occupations", as discussed in the literature, to provide a more comprehensive evaluation of gender biases across the occupational spectrum.

Ethical Considerations

In conducting this work, we acknowledge the ethical responsibility of ensuring our methods accurately detect gender bias to avoid unintentionally contributing to "fairwashing"—the portrayal of biased models as fair. To address this, we intentionally simplified our approach and metrics to maintain transparency in our methodology. This design choice ensures that, even if certain components of our pipeline do not perform as expected, the rationale behind each step remains clear, facilitating easy investigation of any irregularities that could compromise the integrity of our approach. Furthermore, we evaluated our method across a broad range of occupations, recognizing the importance of capturing diverse contexts to provide a more comprehensive and ethically sound analysis of gender bias in machine translation systems.

References

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

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

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

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

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

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

Oisín Deery and Katherine Bailey. 2022. The bias dilemma: the ethics of algorithmic bias in natural-language processing. *Feminist Philosophy Quarterly*, 8(3/4):1–28.

Erik Derner, Sara Sansalvador de la Fuente, Yoan Gutiérrez, Paloma Moreda, and Nuria Oliver. 2024. Leveraging large language models to measure gender bias in gendered languages. *arXiv e-prints*, pages arXiv–2406.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

European Commission. 2020. Gender equality strategy 2020-2025. https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/gender-equality/gender-equality-strategy_en. Accessed: 2025-05-08.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Anna Farkas 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.

Sarthak Garg, Mozhdeh Gheini, Clara Emmanuel, Tatiana Likhomanenko, Qin Gao, and Matthias Paulik. 2024. Generating gender alternatives in machine translation. *arXiv preprint arXiv:2407.20438*.

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

Eleni Gkovedarou, Joke Daems, and Luna De Bruyne. 2025. Gender bias in english-to-greek machine translation. *3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*.

Hila Gonen and Kellie Webster. 2020. Automatically identifying gender issues in machine translation using perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1991–1995, Online. Association for Computational Linguistics.

Atmika Gorti, Aman Chadha, and Manas Gaur. 2024. Unboxing occupational bias: Debiasing llms with us labor data. In *Proceedings of the AAAI Symposium Series*, volume 4, pages 48–55.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Pascal M Gygax, Alan Garnham, and Sam Doehren. 2016. What do true gender ratios and stereotype norms really tell us? *Frontiers in psychology*, 7:1036.

Janiça Hackenbuchner, Joke Daems, and Eleni Gkovedarou. 2025. Genderous: Machine translation and cross-linguistic evaluation of a gender-ambiguous dataset. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 302–319.

Aida Kostikova, Joke Daems, and Todor Lazarov. 2023. How adaptive is adaptive machine translation, really? a gender-neutral language use case. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 95–97, Tampere, Finland. European Association for Machine Translation.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

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

Manuel Lardelli and Dagmar Gromann. 2023. Gender-fair post-editing: A case study beyond the binary. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 251–260, Tampere, Finland. European Association for Machine Translation.

Susan Leavy, Gerardine Meaney, Karen Wade, and Derek Greene. 2020. Mitigating gender bias in machine learning data sets. In *Bias and Social Aspects in Search and Recommendation: First International Workshop, BIAS 2020, Lisbon, Portugal, April 14, Proceedings 1*, pages 12–26. Springer.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *Preprint*, arXiv:2411.16594.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Xianming Li and Jing Li. 2024. AoE: Angle-optimized embeddings for semantic textual similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1825–1839, Bangkok, Thailand. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, pages 189–202.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1393–1409. Association for Computational Linguistics.

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

Orfeas Menis-Mastromichalakis, George Filandrianos, Maria Symeonaki, Glykeria Stamatopoulou, Dimitris Parsanoglou, and Giorgos Stamou. 2025. Gender bias in machine learning: insights from official labour statistics and textual analysis. *Quality & Quantity*, pages 1–35.

Johanna Monti. 2020. Gender issues in machine translation: An unsolved problem? In *The Routledge handbook of translation, feminism and gender*, pages 457–468. Routledge.

Angela Balducci Paolucci, Manuel Lardelli, and Dagmar Gromann. 2023. Gender-fair language in translation: A case study. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 13–23, Tampere, Finland. European Association for Machine Translation.

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

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Silvia Alma Piazzolla, Beatrice Savoldi, and Luisa Bentivogli. 2023. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems. *arXiv preprint arXiv:2306.05882*.

Andrea Piergentili, 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*, pages 71–83, Tampere, Finland. European Association for Machine Translation.

Andrea Piergentili, 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, Singapore. Association for Computational Linguistics.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025. An llm-as-a-judge approach for scalable gender-neutral translation evaluation. *arXiv preprint arXiv:2504.11934*.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Argentina Anna Rescigno, Eva Vanmassenhove, Johanna Monti, Andy Way, et al. 2020. A case study of natural gender phenomena in translation-a comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish. In *CEUR Workshop Proceedings*, pages 359–364. AILC-Associazione Italiana di Linguistica Computazionale.

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

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

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof, and Luisa Bentivogli. 2024. What the harm? quantifying the tangible impact of gender bias in machine translation with a human-centered study. *arXiv preprint arXiv:2410.00545*.

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

Eva H Shinar. 1975. Sexual stereotypes of occupations. *Journal of vocational behavior*, 7(1):99–111.

Jeanette Silveira. 1980. Generic masculine words and thinking. *Women's Studies International Quarterly*, 3(2-3):165–178.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2011. Representation of the sexes in language. In *Social communication*, pages 163–187. Psychology Press.

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

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

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Eva Vanmassenhove. 2024. Gender bias in machine translation and the era of large language models. *Gendered Technology in Translation and Interpreting: Centering Rights in the Development of Language Technology*, page 225.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Eva Vanmassenhove and Johanna Monti. 2021. gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online. Association for Computational Linguistics.

Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and André FT Martins. 2024. Watching the watchers: Exposing gender disparities in machine translation quality estimation. *arXiv preprint arXiv:2410.10995*.

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

A Implementation details

For the generation of the pipeline validation dataset, the benchmarking dataset and the pipeline for extracting occupations along with their genders, we utilized Claude-Sonnet-3.5 v2[8]. The MT systems evaluated in this work are presented in Table 4. For detecting occupations, we applied a cosine similarity threshold of 0.8 when comparing the LLM-generated descriptions with ISCO-08 entries. As described in the main text under our occupation detection procedure, any detected term with a similarity score below this threshold was discarded as a hallucination.

B Prompts

The prompt used by our generation process is presented below.

Generation prompt

Generate a <category> that explicitly mentions the occupation '<occupation title>' in its correct context. Keep it concise. Ensure that no other occupations are mentioned in the text. Ensure the occupation is referred to in a <gender> way, using pronouns, direct mentions, or other linguistic cues.

The category refers to one of the text types, namely short stories, brief news reports, short statements, short conversations, and short presentations. The occupation title is provided exactly as listed in ISCO-08. For example, an occupation is "City Councillor." For the benchmarking dataset (GAMBIT), the <gender> parameter was always set to *Not Clear*, while for the evaluation dataset it was set to either *Masculine* or *Feminine*. Lastly, for text generation in different languages (Greek and

[8] https://openrouter.ai/anthropic/claude-2

MT name	URL
NLLB 600M	https://huggingface.co/facebook/nllb-200-distilled-600M
NLLB 1.3B	https://huggingface.co/facebook/nllb-200-1.3B
M2M100	https://huggingface.co/facebook/m2m100_418M
EuroLLM	https://huggingface.co/utter-project/EuroLLM-1.7B
GT	https://pypi.org/project/googletrans/
Claude	https://openrouter.ai/anthropic/claude-2

Table 4: Machine translation (MT) models used, along with their corresponding hyperlinks.

French) for the evaluation dataset, the phrase "The text should be in <language>." is appended at the end, where <language> is either "Greek" or "French."

For the extraction of the occupation and gender identification, two separate messages were provided in the same chat. The first message identifies the occupation and provides a description that can be used for matching with ISCO's description, while the second message is for gender identification. These messages are presented below. The examples used for few-shot learning were fixed and always provided in the corresponding language (Greek or French), but here we present the prompt with the examples translated into English for clarity.

Message 1

In the following text, identify the occupation titles that are explicitly stated and provide the occupation title along with a brief definition in the following format:

Occupation title: [Occupation title exactly as it is referred to in the text]
Definition: [Definition]
If no occupation is identified, please respond with: "No occupation found."

Here is an example:

Text:
He is a butcher and he is a lawyer.
Occupation title: Butcher
Definition: <definition>
Occupation title: Lawyer
Definition: <definition>

Text:
<text>

Message 2

Please now provide the gender of each identified occupation.
Select from one of the following options:

Masculine if you identified in the text that the occupation refers to a masculine gender.
Feminine if you identified that the occupation refers to a feminine gender.
Not clear if, based on the text, you cannot determine the gender of the occupation.
You must be certain before providing the gender of the occupation and have a clear indication of its gender.

You must answer using only one of the three options and nothing else.

For example:

Text: He is a butcher, and he is a lawyer.
Answer:
Butcher: Masculine
Lawyer: Masculine

Text: <text>
Answer:

For the use of LLMs as translation systems, the prompts used for Claude 3.5 are the following:

Translation Prompt

Translate the following text from English to {target_lang}. Provide only the translated text, without any additional context.
Text:
{source_text}

where target_lang refers to the target language, either Greek or French.

For EuroLLM, we follow the template proposed in the official repository [9].

C GAMBIT

Table 5[10] indicates the average character and word length of each type of text contained in GAMBIT.

[9] https://huggingface.co/utter-project/EuroLLM-1.7B

[10] Tokenization was conducted using https://www.nltk.org/api/nltk.tokenize.word_tokenize.html

Category	$Avg(\|Char\|)$	$Avg(\|Words\|)$
Short story	613.84 ± 78.97	108.13 ± 14.42
Brief news report	537.87 ± 90.35	85.66 ± 14.74
Short statement	132.42 ± 24.96	20.57 ± 3.76
Short conversation	326.2 ± 61.62	70.76 ± 12.16
Short presentation	744.93 ± 143.85	118.43 ± 20.78

Table 5: Average character and word length of samples per category.

Category	$Avg(\|Char\|)$	$Avg(\|Words\|)$
Short story	675.15 ± 71.82	113.44 ± 11.99
Brief news report	546.38 ± 72.48	88.89 ± 11.78
Short statement	147.22 ± 34.42	23.21 ± 5.35
Short conversation	375.7 ± 61.29	73.75 ± 10.79
Short presentation	749.73 ± 110.58	118.92 ± 15.93

Table 6: Average character and word length of samples per category for the French dataset.

D Pipeline Validation Dataset

The character and word length statistics for the pipeline validation datasets are shown in Table 6 and Table 7 for French and Greek, respectively.

In the constructed dataset, each instance comprises not only the textual content but also a set of associated metadata, including the ISCO code for the relevant occupation, its corresponding title and description (sourced from the official ISCO database), and the gender referenced within the text. This annotation enables a systematic evaluation of the pipeline's performance by comparing the ground truth occupation and gender with the occupation and the gender predicted by the model. Specifically, for each instance we first assess the accuracy of occupation identification by verifying whether the occupations predicted by the system match the ground-truth occupation. Subsequently, we compute the accuracy of gender prediction by checking whether the gender assigned by the pipeline aligns with the ground-truth label. In cases where the system failed to detect any occupation in the sentence, the associated gender prediction was automatically considered incorrect.

Category	$Avg(\|Char\|)$	$Avg(\|Words\|)$
Short story	522.56 ± 64.49	88.08 ± 11.2
Brief news report	457.93 ± 49.24	70.89 ± 8.03
Short statement	127.44 ± 28.57	19.41 ± 4.21
Short conversation	341.29 ± 67.84	66.06 ± 13.57
Short presentation	578.84 ± 94.13	87.51 ± 12.82

Table 7: Average character and word length of samples per category for the Greek dataset.

E Per-model Analysis of Extreme Gender Assignments

Table 8 presents the number of ISCO occupations (4-digit level) for which each MT system exhibits either extreme or balanced gender assignments in Greek and French. We define *extreme gendering* as cases where one gender is used in more than 80% of the translations (i.e., $GRAPE_m^{parity} > 0.6$ or $GRAPE_f^{parity} > 0.6$), and *balanced outputs* as those where gender assignments fall within the 30%–70% range. Most systems overwhelmingly favor one gender per occupation, with more than 90% of ISCOs falling in the extreme category for several models. Google Translate and M2M100, for example, produce extreme gendering in over 420 occupations in Greek. By contrast, the large NLLB model (1.3B) shows the most balanced outputs, with over 90 occupations falling within the moderate range for both Greek and French.

F Stereotypes

To quantify the gender stereotyping of occupations, we used ratings from Shinar (1975), which provide perceived gender associations for 129 occupations on a 1–7 scale (1 = most masculine, 7 = most feminine). Each occupation was manually mapped to its closest corresponding 4-digit ISCO category. When multiple occupations mapped to the same ISCO code, we assigned the average of their ratings to that category. If the mapped occupations exhibited substantial variability (i.e., a rating variance ≥ 1.5), the corresponding ISCO code was excluded to ensure consistency. This process yielded 97 unique 4-digit ISCO codes, each associated with a representative rating indicating the degree of gender stereotyping.

We created the stereotypically masculine, feminine, and neutral groups for our study by grouping occupations with stereotyping rating below 2.5, above 5.5, and between 3 and 5 respectively.

G Stereotypical vs. Real-World Correlations

To better understand the factors influencing model behavior, we examined how closely model outputs align with gender stereotypes and real-world labor statistics. Specifically, we computed the correlation between the $GRAPE_f^{parity}$ indicating the model's predicted gender distribution and (i) the stereotype ratings described in Appendix F, and (ii) actual labor market data (female ratio). As real-world

MT	French		Greek	
	Extreme	Moderate	Extreme	Moderate
NLLB-600M	411	15	413	10
NLLB-1.3B	286	91	285	93
M2M100-418M	417	7	424	5
EuroLLM-1.7B	381	30	334	47
GT	321	6	427	5
Claude	387	27	397	23

Table 8: Number of ISCO occupations showing extreme (<20% or >80%) or balanced (30–70%) gender assignments across translation outputs, separated by language.

MT	Real		Stereotype	
	French	Greek	French	Greek
NLLB-600M	0.47	0.47	0.5	0.5
NLLB-1.3B	0.41	0.21	0.41	0.34
M2M100	0.31	0.33	0.34	0.4
EuroLLM	0.65	0.64	0.7	0.74
GT	0.36	0.29	0.42	0.39
Claude	0.63	0.56	0.69	0.64

Table 9: Correlations between model outputs, real data distributions, and stereotype ratings.

statistics are available at the 3-digit ISCO level, we aggregated both model outputs and stereotype ratings accordingly to ensure a fair comparison. For the stereotype ratings, we grouped the 4-digit ISCO codes by their first three digits and calculated the average rating for each group. To maintain consistency, we excluded any groups where the variance among constituent 4-digit occupations was ≥ 1.5. This resulted in 59 unique 3-digit ISCO groups with representative stereotype ratings. We also filtered the real-world data to retain only those 3-digit ISCO groups that appeared in the stereotype set. The detailed correlation results are presented in Table 9. As we can see, across all models and both languages, the correlation with stereotypical ratings is consistently higher than with real-world labor statistics, suggesting a stronger alignment of model behavior with societal stereotypes than actual workforce distributions.

H Cross-lingual and Intra-lingual Correlations

To examine the consistency of gender biases across target languages, we calculated the Pearson correlation coefficient between the gender distributions produced for the two target languages (French and Greek) for each translation model. The resulting scores reflect how similarly each model assigns gendered translations across the two languages.

As shown in Table 10, most models exhibit strong cross-lingual correlations in their gendered translation patterns, with coefficients exceeding 0.70, suggesting largely shared gender biases across the two target languages. The only notable exception is NLLB-1.3B, whose lower correlation score ($r = 0.4775$) aligns with its generally lower gender bias and reduced reliance on masculine defaults (as discussed in Section 5). This may suggest that the model follows a more language-specific strategy for handling gender, rather than relying on shared internal representations.

Table 10: Pearson correlation between gendered translation distributions across French and Greek for each model.

Model	**Cross-lingual Correlation (r)**
NLLB-600M	0.8563
NLLB-1.3B	0.4775
M2M100	0.8524
EuroLLM	0.7482
GT	0.7103
Claude	0.8949

To further illustrate the internal consistency of each model's gendered behavior, Figures 2 and 3 present intra-model correlation heatmaps across models within each language. These visualizations reveal that NLLB-1.3B also shows reduced alignment with other models in both French and Greek, reinforcing the observation that it diverges more significantly from the broader modeling trends.

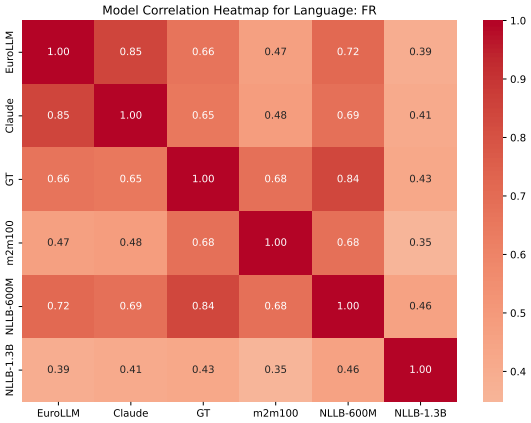

Figure 2: Intra-model correlation of gendered translation distributions in French.

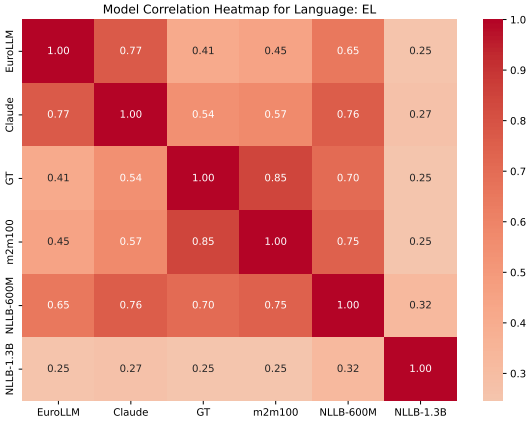

Figure 3: Intra-model correlation of gendered translation distributions in Greek.