**FRI Natural language processing course 2025**
**Project report**

University *of Ljubljana*
Faculty *of Computer and*
*Information Science*

# Investigating Gender Bias in LLM Translations from English to Slovene

Pia Polutnik, Tajda Hladnik, Marko Stoklas

**Abstract**

This project investigates gender bias in machine translation from English to Slovene, focusing on how large language models (LLMs) handle gendered pronouns and professional roles in anti-stereotypical contexts. Using 102 WinoBias sentences translated with ChatGPT, Gemini, and the open-source Helsinki-NLP/OPUS-MT model, we examined gender assignment in isolated translations. We assessed pronoun gender preservation, referent consistency, and grammatical agreement. While all models showed near-balanced pronoun gender distribution, ChatGPT was the most consistent, followed by Gemini, with the open-source Helsinki-NLP/opus-mt-en-sla model showing frequent mismatches. All models exhibited a tendency toward gender stereotypes, especially with ambiguous professions.

**Keywords**

LLMs, machine translation, English-Slovene translation, gender bias, WinoBias, occupational stereotypes

*Advisor: Aleš Žagar*

## Introduction

Large Language Models (LLMs) have in recent years proven to be quite promising with regard to their capabilities in solving NLP (Natural Language Processing) related tasks, including machine translation (MT). One of the unsolved problems remains the occurrence of bias of different types, especially how they handle gendered language during MT. Our project aims to explore this issue further by focusing on MT from English to Slovene. We expect that within this language pair, which differs in the prevalence of gender markers, the LLM must have a certain mean of introducing gender markers. Our project explores gender bias in translation produced by LLMs when translating from English to Slovene, focusing on how occupational roles (e.g. secretary, mechanic) are rendered. he dataset on which we have decided to conduct our research includes the WinoBias dataset, as it offers a sufficient number of sentences which have already been selected due to them expressing gender and occupational bias.

As mentioned, extensive research has been carried out on gender bias in language technologies. The paper ”Investigating Markers and Drivers of Gender Bias in Machine Translations” by P. J. Barclay and A. Sami (2024) shows how automated systems can propagate traditional gender stereotypes, particularly in the context of professions and roles. The authors identify specific linguistic markers, such as gendered pronouns, that often align with stereotypical gender associations in a certain occupations, which can inadvertently perpetuate societal biases.

Similarly, ”Biases in Large Language Models: Origins, Inventory, and Discussion” by R. Navigli, S. Conia, and B. Ross (2024) provides a comprehensive analysis of the origins of bias in large-scale models. The authors argue that these biases are more than a result of training data, but also stem from the way LLMs learn and process language patterns.

Together, these two studies underscore the importance of understanding how LLMs handle gendered pronouns, particularly when translating between languages with different gender structures, and what are the underlying reasons for the occurring biases.

## Methods

Our project examines gender bias in machine translation from English to Slovene, focusing on occupational bias. The research was conducted using the WinoBias dataset, which was accessed through its GitHub repository. A total of 102 sentences (51 sentence pairs) were taken from the *-anti_stereotyped_type2.txt.dev* file of the dataset and used in our research. The file contains pairs of sentences where gender stereotypes are intentionally reversed. These sentences are meant to test whether systems translate gender references correctly even when they go against stereotypical expectations.

Each sentence follows a specific format: *The clerk gave [the mechanic] a present and wished [her] happy birthday*. The sentences contain two individuals in distinct professions, which we designated as referent 1 (the one that appears first) and referent 2 (the individual mentioned second). Referent 1 is not associated with any gendered pronoun, while referent 2 is. Because of this, the gender of referent 2 is always clearly indicated by the pronoun, whereas the gender of referent 1 remains ambiguous in the source sentence.

Firstly, we used OpenAI's ChatGPT and Google's Gemini, two closed-source large language models, to translate the sentences. The prompt for the translation task was: "*Prevedi posamične stavke v slovenščino neodvisno od drugih stavkov.*" ("Translate individual sentences into Slovene independently from the other sentences"). To eliminate any further possibility of the models inferring gender or other contextual information from sentence pairs, we ensured that each sentence in a pair was submitted independently, so that both members of a pair were never processed by the model at the same time.

In addition to closed-source LLMs, we also employed an open-source machine translation model. Specifically, we used the *Helsinki-NLP/opus-mt-en-sla model* from the Hugging Face Model Hub, part of the OPUS-MT collection developed by the Helsinki NLP group. This model is based on the MarianMT framework for translation models. The en-sla model is trained to translate from English several Slavic languages. To ensure the model produced Slovene translations specifically, we prefixed each sentence with the Slovene language tag as required by the multilingual *opus-mt-en-sla* model. The model was accessed in a Google Colab environment using MarianMTModel and MarianTokenizer from the Hugging Face Transformers library. Each of the 102 sentences from the WinoBias dataset was translated individually to prevent any contextual influence between sentence pairs. No additional guidance regarding gender was provided, allowing us to observe the model's default approach to translating gendered pronouns, particularly when occupational roles challenge conventional stereotypes.

The translations were compiled into a Google Docs spreadsheet, where we analyzed the results. We looked at the translations of both referents, and whether the Slovene translation rendered them in masculine or feminine forms. We checked whether the gender of referent 2, explicitly marked by a pronoun in the English source, was preserved correctly in the translation. Since referent 1 is not associated with any gendered pronoun in the English sentence, the model had to decide on the gender by itself. This allowed us to examine how occupations stereotypically linked to a specific gender were translated, noting if the model preserved the anti-stereotypical assignment or reverted to stereotypes. The analysis was conducted by recording how many times each occupation appeared for each of the referents and recording which gender each referent, as well as the pronoun, was translated to.

We continued by recording changes in pronoun gender distribution by comparing the number of masculine and feminine pronouns in the English source sentences (51 each) with the number of each found in the Slovene translations for each model. This allowed us to track whether the models maintained the original gender balance or introduced a shift during translation. To evaluate gender consistency in translation and its subsequent accuracy in our case, we assessed each sentence across three criteria. Each criterion was marked as either correct (T) or incorrect (F) to identify where inconsistencies occurred. Firstly, we checked internal agreement between the pronoun and its referent within the target sentence to evaluate grammaticality (Int. P-P). This was marked as correct if the gender of the pronoun matched the grammatical gender of its referent within the Slovene sentence, and incorrect if they did not align. Secondly, we examined whether the gender of the pronoun matched across the source and target sentences (CS. P-P). Lastly, we checked whether the intended gender of referent 2 in English was preserved in the translation (CS. R-R). A mismatch in either of the first two criteria resulted in a mismatch for the third, which captures the overall accuracy of gender representation.

## Figures

The following figures showcase, firstly, the change in configuration of the number of pronouns pertaining to a certain gender, and secondly the various match criteria we examined across sentence translations. Other figures that regard occupations are found within our Google Sheets folder accessible through our GitHub repository.

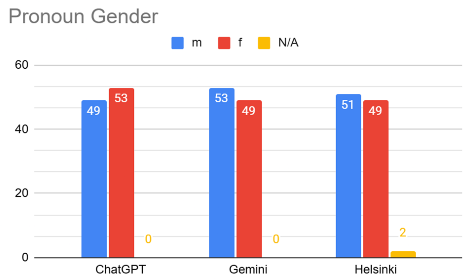

**Figure 1.** Pronoun gender in target sentences across the LLMs. The original division was 51 masculine pronouns and 51 feminine pronouns.

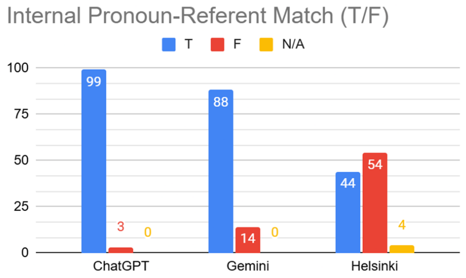

**Figure 2.** Internal Pronoun-Referent Match across LLMs. This figure shows the grammaticality of translations, namely, if the gender of the referent matcher the gender of the pronoun in Slovene.

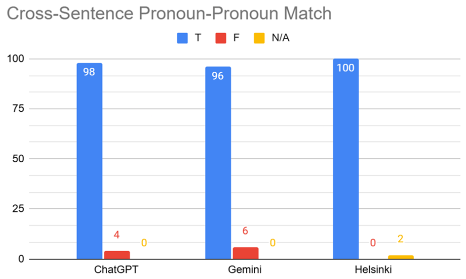

**Figure 3.** Pronoun-Pronoun matches across translations by the LLMs. The figure shows whether the genders of the pronouns are matching across the source and target sentences.

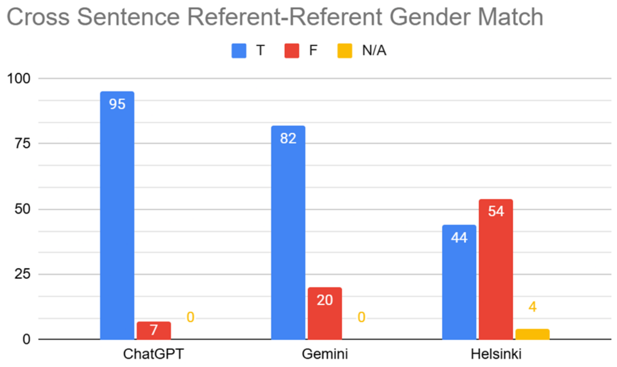

**Figure 4.** Cross Sentence Referent-Referet Gender Match across LLMs. The figure indicates the overall accuracy of translation.

# 1. Results

This section is divided into 2 parts. Firstly, it showcases results pertaining to the tendencies of gendering across sentence translations and the matching of pronouns and referents. And secondly, the role of professional roles and their gender bias.

## 1.1 Tendency of gendering and translation accuracy

The models translated the English sentences into Slovene with varying distributions of gendered pronouns shown in *figure 1*. ChatGPT produced 49 masculine and 53 feminine pronouns, corresponding to 48% and 52%, respectively. Gemini produced 53 masculine and 49 feminine pronouns, while the Helsinki model produced 51 masculine and 49 feminine pronouns. Two sentences from the Helsinki model were excluded from gender classification due to structural changes that replaced or removed pronouns.

Regarding accuracy and consistency of translation taking into account the three criteria mentioned in Methodology, we summarize the following results showcased in *figures 2, 3 and 4*. For ChatGPT, the *Int. P-P* resulted in 99 True and 3 False matches. The *CS. P-P* showed 98 True and 4 False matches. The *CS. R-R* yielded 95 True and 7 False matches. Gemini produced 88 True and 14 False matches for *Int. P-P*, 96 True and 6 False for the *CS. P-P*, and 82 True and 20 False matches for the *CS R-R*. Helsinki-NLP showed 44 True and 54 False for the *Internal pronoun–referent match*, all cases of *CS P-P, except for 2 which were unusable, were True. As for the CS R-R matches,* 44 cases exhibited True , 54 False, and 4 N/A.

## 1.2 Gender bias in occupations

### 1.2.1 ChatGPT

ChatGPT produced 7 inadequate translations. With regards to the Int. P-R matches, there were three sentences where the profession was masculine, but the gender was feminine. These professions included *razvijalec*, *mehanik*, and *kmet*. An example of the sentence includes "*Oblikovalec se je prepiral z razvijalcem in jo udaril po obrazu*." As for the CS P-P, there were four instances where the gender of the referent was changed in translation. Three of these professions were feminine and included *gospodinja* (2) and *prodajalka*. For example: "*The mover said thank you to the housekeeper and give him a hug*" versus "*Selilec se je zahvalil gospodinji in jo objel*." The fourth included a masculine profession, which was translated as *nadzornik* (supervisor). As for the profession of referent 1, ChatGPT produced 32 different masculine forms for professions, and only 15 feminine forms. Out of the 32 masculine forms, the five most common were *oblikovalec* (7), *razvijalec* (5), *mehanik* (5), *receptor* (4), and *kmet* (4). The five most common feminine forms were *frizerka* (4), *čistilka* (3), *gospodinja* (2), *sekretarka* (2), and *urednica* (2).

### 1.2.2 Gemini

Gemini produced 14 inadequate translations. With regards to the Int. P–R matches, there were multiple sentences where the profession was masculine, but the gender marking was feminine. These professions included *razvijalec, mehanik* (5), *selivec, gospodinja, šef, medicinska sestra, gradbeni delavec, analitik*, and *mizar* (2). As for the CS P–P category, there were six instances where the gender of the referent changed in translation. These included *gospodinja, analitik, recepcijonistka, razvijalec, kuhar*, and *varnostnik*. Regarding the profession of referent 1, Gemini produced 31 different masculine forms and 14 feminine forms. Among the masculine forms, the five most common were *razvijalec* (6), *mehanik* (5), *prodajalec* (4), *uslužbenec* (3), and *selivec* (3). Among the feminine forms, the five most frequent were *oblikovalka* (6), *frizerka* (6), *tajnica* (3), *blagajničarka* (3), and *gospodinja* (2).

#### 1.2.3 Helsinki-NLP/opus-mt-en-sla model

The Helsinki-NLP/opus-mt-en-sla model, produced 54 mismatches in the internal pronoun–referent category (*Int. P-R*). The mismatched professions included *razvijalec* (4), *mehanik* (4), *gospodinja* (2), *premikalec, frizerka* (3), *odvetnik* (3), *kuhar* (2), *recepcionarka* (3), *nadzornik, kmet* (2), *preselitelj* (2), *tajnica* (2), *blagajnica, voznik, stražar, upravnik, zdravnik* (3), *delavec* (2), *direktor, menedžer, medicinska sestra, gradbeni delavec* (2), *tesar* (2), *prodajalec, pekar*, and *hišnik*. All CS P–P matches for Helsinki-NLP model were correctly preserved, with no mismatches recorded. Helsinki-NLP produced 36 different masculine and 10 feminine forms for referent 1 professions. Among the masculine forms, the most common were *razvijalec* (6), *mehanik* (4), *kmet* (4), *prodajalec*(4), and *čistilec* (4). For the feminine forms, the most frequent were *frizerka* (5), *oblikovalka* (5), *recepcionarka* (4), *tajnica* (3), and *hišna pomočnica* (1).

## 2. Discussion

### 2.1 Tendency of gendering and translation accuracy

Firstly, the distribution of gendered pronouns across the three models reveals subtle differences in their gender assignment tendencies. While all three models produced a relatively balanced ratio of masculine to feminine pronouns, ChatGPT slightly favored feminine forms, whereas Gemini and Helsinki-NLP/opus-mt-en-sla model leaned toward masculine usage. This may point to the differences in models' architecture and training data, which may have influenced gender preferences. Overall, the variations point to model-specific biases in gender realization during translation, not that masculine forms are preferred a priori - we propose this as an area of further inquiry.

The evaluation of results in terms of *matching* shows that ChatGPT maintained the highest consistency in the internal pronoun–referent match, with only minor mismatches. Gemini performed moderately well, while Helsinki-NLP produced more mismatches than correct matches, indicating limited grammatical reliability. This criteria reflects the grammaticality of translated sentences, however, consistency in this category does not necessarily reflect accuracy in gender translation, as models may prioritize syntactic correctness over being true to the source text. This was especially true for the open-source model, where it seems that it translated the pronoun regardless of the gender of the referent in the translated sentence. The two closed-source models proved more proficient in this regard, though they still had some problems.

In the cross-sentence pronoun–pronoun match, all three models preserved the pronoun's gender consistently, with similar accuracy rates. This suggests that models can translate pronouns reliably across sentences. This may occur independently of referent interpretation, however, and points to a tendency of the language models to fixate on translating the pronoun correctly at the expense of inaccurate translation.

The cross-sentence referent–referent match proved more challenging, with ChatGPT providing the most consistent results, followed by Gemini. Helsinki-NLP/opus-mt-en-sla model underperformed in both grammatical agreement and gender preservation. These outcomes indicate that while pronoun translation is relatively robust across models, maintaining consistent gender for referents across sentence pairs is more difficult.

### 2.2 Gender bias in occupations

#### 2.2.1 ChatGPT

ChatGPT produced 7 inadequate translations. These mismatches point to a tendency of ChatGPT, when making mistakes, to favor masculine forms of professions such as *farmer, mechanic, and developer,* in the case of Int. P-R matches, prioritizing masculine forms of these professions. Furthermore, in the case of CS P-P matches, three of the professions were feminine – housekeeper or clerk, while one was masculine - supervisor. The inconsistent matching between *razvijalec, mehanik and kmet,* showcase that the model, in this case, did not find suitable alternatives for these stereotypically masculine roles. Though this was not the case throughout the entirety of the sentences, since in other cases the model used suitable substitutes. The fact that our team split the sentences and translated them in three batches may have been played a part in this.

As for the model's translation of referent 1, we can see that in the discrepancy between the number of masculine and feminine forms the model is far more creative with naming male professions. The five that were most common include *oblikovalec, razvijalec, mehanik, receptor, kmet.* This does not necessarily mean that it was biased as *oblikovalec (designer) and receptor (receptionist)* are considered feminine occupations in the data set. Still the other three remaining occupations, perhaps unsurprisingly, coincide with the False *Int. P-R* matches, indicating that these occupations experience the greatest gender bias. As for the feminine forms that were most common as referent 1, all of their English counterparts exhibit feminine bias in the data set, and were also translated as feminine. The results overall point that ChatGPT prefers to use masculine forms as the default, although it does show bias when it comes to stereotypically feminine professions. Further inquiry is required to assess whether this is due to Slovene considering the masculine form as default, the male-oriented data on which the model may have been trained (and to what these two intersect), or if there is some other underlying cause.

#### 2.2.2 Gemini

Gemini incorrectly translated 20 sentences, suggesting a higher inconsistency in managing gender agreement, especially when resolving grammatical gender across sentence elements when compared to ChatGPT. The mismatches in Int. P-R category show the model's tendency to incorrectly assign gender when the gender of the referent was clear. Compared to ChatGPT, the LLM in question shows a greater degree of incorrectly integrating a broader sentence context when selecting appropriate grammatical forms in Slovene.

In the cross-sentence analysis, while Gemini did retain pronoun gender in most cases, the observed mismatches points to an occasional failure to track the referent's gender consistently across sentences. It is largely comparable to ChatGPT, with 2 additional mismatches within CS P-P.

There was also a notable imbalance between the number of masculine and feminine profession forms used for referent 1 which reflects an implicit bias toward masculine defaults and is likewise comparable with ChatGPT, with a minimal difference of one profession regarding both masculine and feminine forms. While the model does produce stereotypically feminine roles in some cases, it appears to rely on masculine forms more readily, which reinforces a broader trend that was also evident with ChatGPT.

#### 2.2.3 Helsinki-NLP/opus-mt-en-sla model

Helsinki-NLP/opus-mt-en-sla model demonstrated the weakest performance among the models in terms of internal consistency which regarding the referent and pronoun. The high number of mismatches in pronoun–referent agreement suggests the model struggles to align grammatical structure with semantic context in Slovene, particularly with regards to gender markers. This showcases its limitations as a translation tool. This may be accounted for by its training data or lower capacity to manage contextual dependencies.

This was also evident in its performance in the cross-sentence pronoun–pronoun category where Helsinki-NLP/opus-mt-en-sla model was perfectly consistent. This may indicate that the LLM would always translate the constituents of the sentence one to one, disregarding the context, hence why it may have translated the professions into the masculine (even though the feminine form was intended) but correctly translated the pronoun. This may also be the reasoning behind the previous two models making the same, albeit drastically fewer, mistakes.

In terms of occupational bias derived from analyzing the profession and gender of referent 1, Helsinki-NLP heavily favors masculine forms, displaying the most skewed distribution of the three models. Its minimal use of feminine forms may stem from both its reliance on masculine form as a standard in the Slovene language and potentially a narrower, and therefore more biased training corpus.

#### 2.2.4 Comparative Discussion

Across all three models we can observe that stereotypes continue to appear, although in varying degrees. ChatGPT demonstrated the most balanced and consistent output, both in grammatical agreement and referent gender tracking. Gemini followed with moderate performance, and Helsinki-NLP/opus-mt-en-sla model significantly lagged behind, particularly in internal grammatical agreement and its incessant Cross-Sentence Pronoun-Pronoun matching. While all models performed well in preserving pronoun gender across sentences, this success appears very mechanical—consistent, but not necessarily correct as the results suggest. This was especially evident in Helsinki-NLP/opus-mt-en-sla model's case, where consistency did not translate to accuracy. The disparities suggest that larger, more advanced models like ChatGPT are better at capturing nuanced contextual dependencies required for correct gender assignment in translation, and the same holds true for Gemini, although ChatGPT offers a far more accurate translation service. The Helsinki-NLP/opus-mt-en-sla model is far more susceptible to surface-level patterns or stereotypical associations that the latter two. Furthermore, it may be the case that models retained gender forms from earlier sentences without aligning them with the referent's intended identity in the sentence they tried to translate later on - clinging onto context from previous sentences. This may also have played the part in our research. Still, regarding gender bias and the question whether LLMs are equipped with the necessary means of tackling issues that pertain to gendered language, the answer remains unclear. The closed-source model is largely inconsistent, which is also evident in the general quality of translations. While the two open-source par far better with the task of translation they nevertheless exhibit some degree of error. While the language models may be at fault in this regard - alongside their training data, this may reflect on our state of language, and should encourage us to produce more variable, less biased data sets in order to improve LLMs and facilitate visibility and nuance.

## References

Barclay, P. J., & Sami, A. (2024). Investigating markers and drivers of gender bias in machine translations. *arXiv*. https://doi.org/10.48550/arxiv.2403.11896

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. https://dl.acm.org/doi/abs/10.1145/3582269.3615599

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality, 15*(2), Article 10. https://doi.org/10.1145/3597307

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT — Building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)* (pp. 479–480). European Association for Machine Translation. https://huggingface.co/Helsinki-NLP/opus-mt-en-sla

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conferences of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018), Volume 2 (Short Papers)* (pp. 15–20). https://doi.org/10.18653/v1/N18-2003